Alimentation automatique d'une base de connaissances à partir de textes en langue naturelle. Application au domaine de l'innovation

Issam Al Haj Hasan

N° D'ORDRE : D.U. 1879
EDSPIC : 419

Universite Blaise Pascal – Clermont-Ferrand II

ECOLE DOCTORALE
SCIENCES POUR L'INGENIEUR DE CLERMONT-FERRAND

THESE

Présentée par

Issam AL HAJ HASAN

DEA : Informatique, productique, immagerie médicale

Pour obtenir le grade de

DOCTEUR D'UNIVERSITE

Spéciatié : INFORMATIQUE

Alimentation automatique d'une base de connaissances à partir de textes en langue naturelle

Application au domaine de l'innovation

Soutenue publiquement le 20 novembre 2008 devant le jury :

Rapporteurs :	Pr. André FLORY
	Pr. Chantal SOULE-DUPUY
Examinateurs :	Pr. Grigore GOGU (Co-Directeur de thèse)
	Pr. Jérôme DARMONT
	Pr. Michel SCHNEIDER (Co-Directeur de thèse)
	Pr. Farouk TOUMANI

Table des matières

Liste des tabeaux

Table des Figures

Introduction

Dans de nombreux domaines apparaît le besoin de construire des bases de connaissances à partir de données diverses : bases de données, documents internes ou externes plus ou moins structurés. Très souvent le repérage et la saisie de ces données et connaissances sont effectués manuellement. A cause du nombre croissant de ressources disponibles sur le Web, ces processus manuels doivent être automatisés partiellement ou totalement. L'extraction automatique d'informations à partir de textes en langage naturel a fait l'objet de très nombreux travaux au cours de ces dernières années. Des résultats significatifs ont été obtenus mais plusieurs problèmes restent ouverts.

Les systèmes d'extraction sont basés sur des règles d'extraction utilisant des relations syntaxiques et sémantiques dans le texte. La spécification de telles relations dans les textes en langage naturel ou semi structurés est très difficile et demande beaucoup d'expérience. C'est pourquoi la plupart des systèmes d'extraction implémentés utilisent l'apprentissage. Ces systèmes visent à automatiser la production des règles d'extraction et à faciliter l'adaptation à un nouveau besoin ou à une nouvelle application. Le rôle de l'utilisateur dans ce cas est d'étiqueter des textes pour entraîner le système qui en déduit la formulation des règles d'extraction. La difficulté pour ces systèmes est de savoir choisir les textes, de savoir les étiqueter et de savoir terminer la phase d'apprentissage.

Dans ce travail nous nous intéressons à l'alimentation automatique d'une base de connaissances pour l'innovation. L'innovation est une activité qui s'intéresse aux théories et aux expérimentations pour résoudre un problème de conception ou de production. C'est une activité multi domaine qui n'est pas spécifique à une filière ou à une autre. Les principaux travaux de recherche relatifs à l'innovation ont surtout concerné l'innovation technique. Plusieurs outils et bases de connaissances ont déjà été élaborés pour aider à la résolution créative des problèmes, en se basant sur la méthode d'innovation TRIZ. Ces outils et leurs bases de connaissances permettent à l'utilisateur d'analyser, évaluer et optimiser son problème et sa solution.

Cependant, ces bases de connaissances doivent être mises à jour en permanence. De nouveaux exemples doivent être introduits régulièrement. De nouvelles opérations et méthodes d'innovation doivent être définies par les compagnies productrices en fonction des progrès des sciences et des technologies. Des systèmes d'aide à l'innovation adaptables à chaque domaine et permettant à un expert de repérer semi-automatiquement des solutions pertinentes à son problème deviennent de plus en plus nécessaires.

Ainsi, notre projet vise à concevoir un système d'aide à l'innovation plus facile à exploiter et à adapter aux besoins. Ce système doit permettre deux tâches complémentaires : la tâche d'enrichissement automatique d'une base de connaissances par des exemples de résolution inventive des problèmes et la tâche de résolution semi-automatique des problèmes posés par un expert.

En nous basant sur des concepts définis dans la méthode d'innovation TRIZ, nous avons conçu une ontologie d'innovation. Cette ontologie est une représentation sémantique d'une base de connaissances. Elle permet à l'expert de définir des opérateurs d'innovation. Un opérateur d'innovation est un concept général auquel on doit associer des exemples et sur lequel on se base pour la résolution des problèmes.

Pour concevoir un module d'enrichissement automatique de la base de connaissances par des exemples, nous nous sommes appuyés sur les systèmes d'extraction d'information et leurs techniques. Nous avons constaté que ces systèmes ne sont pas facilement adaptables à notre application. Dans les systèmes fondés sur une approche linguistique la mise au point des règles d'extraction est très compliquée. Dans les systèmes fondés sur l'apprentissage, l'élaboration d'un corpus d'entraînement d'une taille suffisante pour chaque opérateur d'innovation et très difficile.

Par conséquent, nous nous sommes focalisés dans cette thèse sur l'élaboration d'une approche d'extraction pour implémenter le module d'enrichissement automatique. Le module d'aide à la résolution de problèmes n'a pas été investigué. Nous avons simplement conçu et implémenté une interface graphique qui permet à l'utilisateur de rechercher les exemples alimentés dans la base de connaissance et les textes associés.

Dans le développement de ce module, nous avons cherché à minimiser les ressources de connaissances nécessaires à l'extraction afin de faciliter son adaptation à tout domaine. Ces

ressources sont constituées par la base de connaissances, l'ontologie et un mapping entre des entités de la base vers des termes dans les textes.

Le mémoire est organisé comme suit. Le domaine de l'innovation et ses concepts sont étudiés et présentés dans le chapitre 1. L'état de l'art sur les systèmes d'extraction, les systèmes de question-réponse et leurs approches est étudié et présenté dans le chapitre 2. L'architecture de notre système d'aide à l'innovation et les problèmes rencontrés pour sa mise en œuvre sont présentés dans le chapitre 3. Dans le chapitre 4, nous présentons notre ontologie d'innovation O_{ino}, des opérateurs définis dans une base de connaissances BC et le mapping *Map* à des termes. Dans le même chapitre, nous présentons ensuite une approche d'extraction d'exemples à partir des textes et une approche pour le repérage des textes sur le Web. Dans le chapitre 5, nous présentons l'implémentation du prototype et une évaluation de notre proposition à partir de différentes expérimentations. Nous terminons le mémoire par une conclusion et quelques perspectives.

Chapitre 1

Domaine de l'innovation et problématique étudiée

1.1 Domaine de l'innovation

L'innovation ou la créativité est l'acte d'introduire quelque chose de nouveau. Cette notion globale peut être de diverses natures (produit, processus, méthode, organisation, etc.). Elle concerne tous les domaines (scientifiques, technologiques, économiques, artistiques, etc.). Dans les entreprises, où la compétition technico-sociale [8] incite à améliorer le prix, la qualité et la complexité des produits, l'innovation représente bien souvent un enjeu de survie.

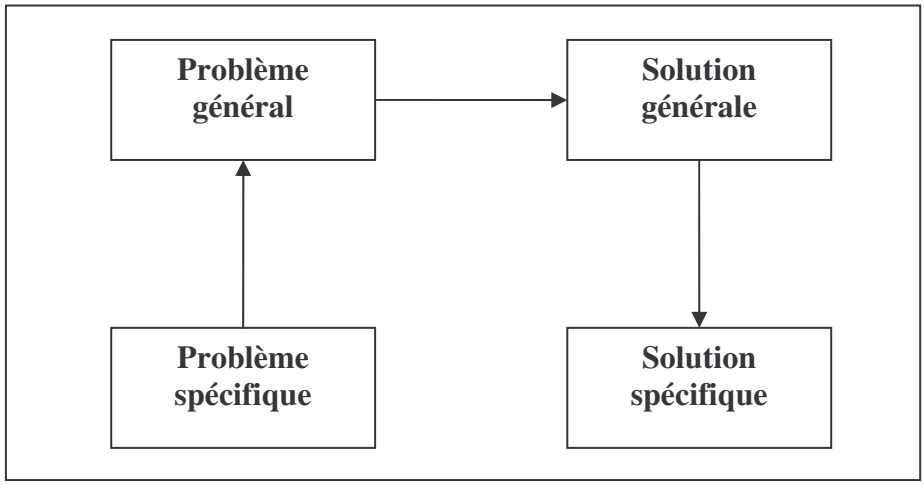


Figure 1.1: Processus de conceptualisation mis en œuvre dans TRIZ

Des approches pour la gestion de connaissances spécifiques comme TRIZ [1 à 10], WOIS « *Contradiction Oriented Innovation Strategy* » [11, 12], USIT « *Unified Structured Inventive Thinking*» [13], QFD « *Quality Function Deployement* » [7, 14, 15], CPS « *Creative Problem Solving* » [16, 17], etc. ont été proposées pour systématiser l'innovation technique et pour mobiliser les connaissances des entreprises. Ainsi, ces entreprises peuvent résoudre leurs problèmes sans faire appel aux spécialistes extérieurs, à chaque fois qu'un problème apparaît.

Sur le marché, la méthode TRIZ demeure la seule méthode à aller jusqu'à proposer des idées et des pistes technologiques de recherche [8]. Les autres méthodes d'innovation viennent en complément ou s'en inspirent. TRIZ permet d'explorer systématiquement le domaine des solutions possibles à un problème donné, y compris d'exploiter des solutions similaires appliquées à d'autres domaines, puis de générer des concepts innovants [9]. A partir d'un problème spécifique (cf. figure 1.1), il s'agit dans un premier temps de formuler un problème général (ou standard), puis d'utiliser TRIZ pour déterminer les solutions génériques, et enfin d'interpréter ces solutions génériques pour en tirer des solutions spécifiques au problème.

Cette méthode a été initialement développée dans ce qui fut l'Union Soviétique et TRIZ est l'acronyme russe pour « *Téoria Rechénia Izobrétatelskikh Zadatch* ». Reprise ensuite aux Etats Unies sous le vocable TIPS « *Theory of Inventive Problem Solving* », elle commence ces dernières années à se diffuser largement dans le monde.

Elle a été mise au point par Genrich Saulovich ALTSHULLER (1926-1998). A partir de 1946, Altshuller a commencé à inspecter des brevets et des documents techniques pour extraire des règles de résolution de problèmes aidant les autres à innover. Ce travail s'est poursuivi durant de nombreuses années pendant lesquelles plus de 2 millions de brevets ont été analysés. Les principaux concepts issus de ce travail et définis dans la méthode TRIZ sont présentés dans la section suivante.

Dès l'origine, Altshuller s'est appuyé sur de nouveaux critères pour analyser les brevets. En fonction de la nature des connaissances à mobiliser et du nombre de solutions à considérer, appelé nombre d'essais et erreurs, les brevets sont attribués à un niveau d'innovation. Cinq niveaux d'innovation sont définis, de la solution apparente jusqu'à la découverte (cf. tableau 1.1). Les brevets représentant une petite modification dans la conception sont attribués au plus bas niveau. Les brevets qui changent plus profondément le système sont considérés plus inventifs et ceux qui introduisent une nouveauté scientifique sont considérés comme les plus importants [4].

A partir de cette classification, Altshuller a constaté que plus de 90% des problèmes rencontrés ont été résolus quelque part avant. Si les ingénieurs pouvaient suivre un cheminement vers une solution idéale, en commençant par le niveau le plus bas en monopolisant leurs connaissances personnelles et leur expérience, puis en procédant par

induction vers des niveaux plus élevés, la plupart des solutions pourraient être tirées de la connaissance déjà présente dans l'entreprise, l'industrie ou dans une autre industrie [2, 4, 6, 7].

Niveau	**Degré d'inventivité**	**% de solutions**	**Origine des connaissances**	**Nombre approx. de solutions à considérer**
1	Solution apparente	32	Connaissances d'un individu	10
2	Amélioration mineure	54	Connaissances de l'entreprise	100
3	Amélioration majeure	18	Connaissances de l'industrie	1.000
4	Nouveau concept	4	Connaissances de toutes les industries confondues	100.000
5	Découverte	1	Ensemble des savoirs	1.000.000

Tableau 1.1: Degrés d'inventivité

1.2 Concepts de la méthode d'innovation TRIZ

Dans cette section, on présente les concepts définis et utilisés dans la méthode TRIZ. Ces concepts représentent, en fait, une terminologie d'une base de connaissances d'innovation. Cette terminologie permet à l'utilisateur de la base ou de la méthode TRIZ (expert, ingénieur, étudiant, etc.) de repérer ou d'intégrer une information dans la base.

De plus, une représentation sémantique de cette terminologie, nous paraît nécessaire pour automatiser totalement ou partiellement les applications du domaine de l'innovation comme la recherche documentaire, l'extraction d'informations, la conception et la formulation de problèmes.

1.2.1 Système technique

Un système technique est un ensemble de composants (cf. § 1.2.2) associés et organisés par un modèle de fonctionnement (cf. § 1.2.4) pour donner un produit exigé (cf. § 1.2.6). La figure 1.2 présente un système technique : la seringue et ses composants.

1.2.2 Composant

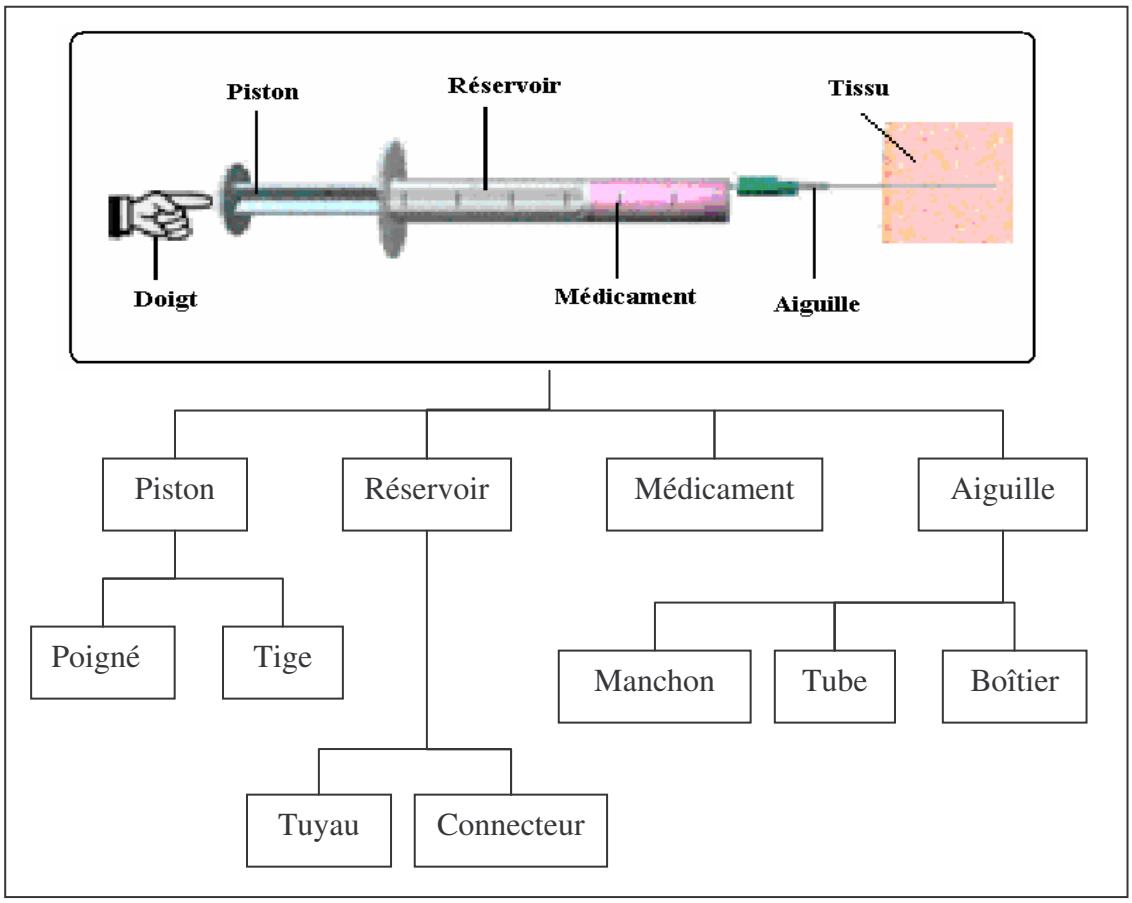


Figure 1.2: Système technique et ses composants

Un composant est une partie constitutive d'un système technique [22]. Cette partie est nécessaire pour compléter ou améliorer une fonctionnalité dans le système. Un composant peut être une ressource ou un sous-système possédant ses propres composants.

1.2.3 Ressource

C'est un élément disponible dans l'environnement et nécessaire pour compléter ou améliorer un système et le déplacer vers l'idéal (cf. § 1.2.9). Cet élément peut être naturel ou dérivé d'un élément naturel (après avoir subi un traitement). Ces ressources sont classées dans six groupes [7] :

1. Substance : déchet, matière primaire, produit d'un système (cf. § 1.2.6)… ;
2. Champ de force et énergie ;

3. Espace : l'espace exigé peut être vide (i.e. afin d'augmenter la durée de la vie d'un disque dur, on le met dans une espace vide) ; il a une forme et une dimension prédéfinies ; il peut être vertical, imbriqué… ;

4. Temps : le temps s'exprime par rapport à un processus (ou travail), il est planifié dans le processus, il peut se rapporter à un pré-processus ou à un post-processus ;

5. Information (mobile, transitoire, changement d'état, propriété héritée);

6. Fonction (primaire d'une ressource, secondaire, nuisible).

Un des moyens permettant de généraliser une solution dans TRIZ est de regrouper sémantiquement les ressources employées dans les brevets, en se basant sur des propriétés communes. On peut par exemple envisager les regroupements suivants : solide, gaz, plasma, charge, force [24, 25] (cf. annexe C).

1.2.4 Modèle de fonctionnement

Il décrit les fonctions agissant entre les composants d'un système [22] (cf. figure 1.3). Au plus haut niveau opèrent les fonctions du super-système (cf. § 1.2.7) et au plus bas niveau on trouve le produit (cf. § 1.2.6).

1.2.5 Fonction

Une fonction est une relation Action-Objet *(A-O)* décrivant dans un système le rôle principal (Action) d'un composant (Sujet) sur un autre composant ou un paramètre (Objet) [7, 21, 22]. Par exemple, dans le système présenté dans la figure 1.3, le piston a la fonction (*Déplace-Médicament*).

La réalisation d'une fonction entraîne souvent des effets. Par exemple le déplacement de médicament change son volume dans le réservoir. Une grande majorité de fonctions sont réalisées par des effets (cf. § 1.2.13.3). Par exemple, on utilise un champ magnétique et une boucle pour produire un champ électrique, mais pour cela il faut alterner le champ magnétique dans la boucle. Une fonction produite peut être bénéfique (désirable) ou nuisible (indésirable). Par exemple, dans la même figure 1.3, l'insertion de l'aiguille dans le tissu l'endommage : (*insertion-aiguille*) → (*endommage-tissu*).

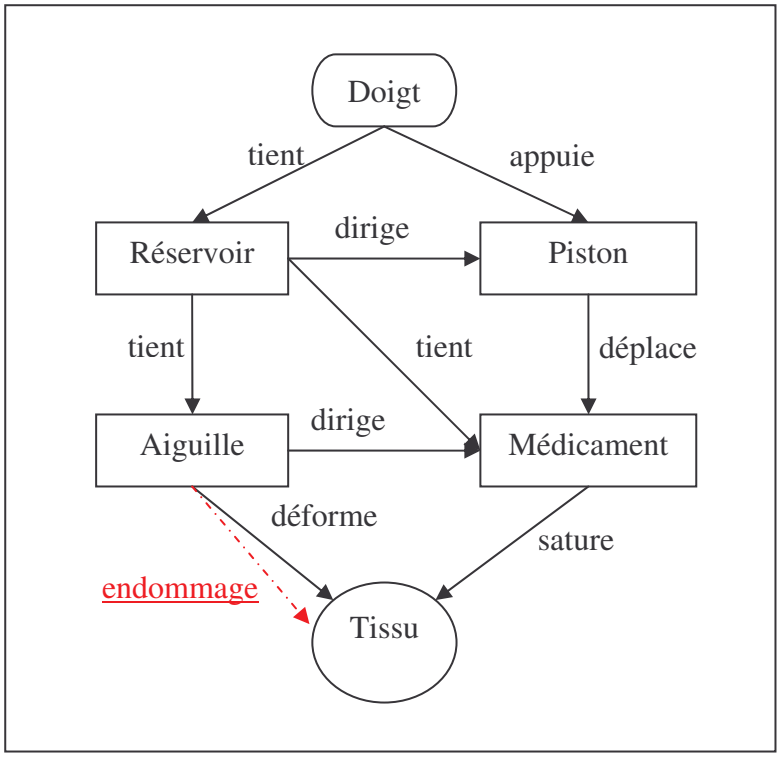


Figure 1.3: Modèle de fonctionnement

1.2.6 Produit

C'est le résultat principal d'un système et il n'en fait pas partie [7, 22]. Il peut être simplement une ressource (substance, énergie, information, fonction, etc.) créée ou modifiée par le système. Par exemple :

1. Un capteur génère un signal. Le signal est un produit car il est produit par le capteur.
2. Une pompe applique une pression sur l'eau. L'eau est le produit car elle est modifiée par la pompe.
3. Dans la figure 1.3, le produit est le tissu modifié par l'injection du médicament.

1.2.7 Super-système

C'est un élément de l'environnement du système interagissant avec lui (ayant des fonctions sur le système) et il n'en fait pas partie [7, 21, 22]. C'est par exemple le doigt dans la figure 1.3.

1.2.8 Sous-système

C'est un système technique qui fait partie (composant) d'un autre système ayant ses propres composants et son propre produit (cf. § 1.2.6).

1.2.9 Idéalité

C'est une notion qui exprime le résultat idéal (optimal) vers lequel doit tendre le concepteur. Le résultat ultime idéal serait celui pour lequel tous les effets utiles du fonctionnement d'un système sont assurés et tous les inconvénients (consommation d'énergie, danger, pollution,…) éliminés. L'idéalité est mathématiquement définie par la somme des effets utiles E_u d'un système S sur la somme des effets indésirables E_i [6].

$$idéalité = \frac{\sum_{E_u \in S} E_u}{\sum_{E_i \in S} E_i}$$

1.2.10 Problème d'innovation

C'est un problème contenant des contradictions et le passage à la solution est inconnu [2, 4]. Par exemple dans la figure 1.2, comment peut-on insérer le médicament dans le tissu sans l'endommager?

1.2.11 Contradiction

Altshuller a défini deux types de contradiction [2, 6]: la contradiction technique et la contradiction physique.

1.2.11.1 Contradiction technique

Dans ce type de contradiction, une solution évidente pour améliorer une caractéristique (cf. § 1.2.12) utile *A* conduit à la dégradation d'une autre caractéristique utile *B*. Par exemple, pour livrer des pizzas chaudes aux clients, on peut augmenter l'épaisseur du carton. Mais cette solution conduit à accroître la masse de la boîte à pizza, et donc à consommer davantage de carton. Ce type de contradiction est résolu en utilisant la matrice de résolution et les 40 principes inventifs (cf. § 1.2.13.1).

1.2.11.2 Contradiction physique

Ce type de contradiction existe quand le bon fonctionnement d'un système exige des caractéristiques contradictoires comme : forte et faible, dur et mou, lisse et rugueux. Par exemple, l'avion doit avoir une très grande portance en plein vol et une très grande trainée pendant la phase d'atterrissage. Ce type de contradiction est résolu en utilisant des principes de séparation (cf. § 1.2.13.2).

1.2.12 Paramètre (caractéristique) d'innovation

N.	Paramètre (caractéristique)	N.	Paramètre (caractéristique)
1	Masse / poids d'un objet mobile	2	Masse / poids d'un objet fixe
3	Longueur d'un objet mobile	4	Longueur d'un objet fixe
5	Surface d'un objet mobile	6	Surface d'un objet fixe
7	Volume d'un objet mobile	8	Volume d'un objet fixe
9	Vitesse	10	Force
11	Contrainte ou pression	12	Forme
13	Stabilité de la composition d'un objet	14	Résistance
15	Durée de l'action d'un objet mobile	16	Durée de l'action d'un objet fixe
17	Température	18	Brillance
19	Utilisation d'énergie d'un objet mobile	20	Utilisation d'énergie d'un objet fixe
21	Puissance	22	Perte d'énergie
23	Perte de substance	24	Perte d'information
25	Perte de temps	26	Quantité de substance
27	Fiabilité	28	Précision de la mesure
29	Précision de l'usinage	30	Facteurs néfastes agissant sur l'objet
31	Facteurs néfastes générés par l'objet	32	Usinabilité
33	Facilité d'utilisation	34	Facilité de réparation
35	Adaptabilité	36	Complexité de l'appareil
37	Difficultés de détection et de mesure	38	Degré d'automatisation
39	Productivité		

Tableau 1.2: Paramètres d'innovation extraits par Altshuller

A partir de 40.000 brevets d'invention étudiés, Altshuller a extrait 39 caractéristiques (paramètres) qui génèrent des contradictions techniques (cf. tableau 1.2).

1.2.13 Opérateur d'innovation

Un opérateur d'innovation est un concept général recommandé pour résoudre un problème apparu dans un système. La méthode TRIZ définit quatre groupes d'opérateurs. Ce sont les principes inventifs, les principes de séparation, les effets scientifiques et les solutions innovantes génériques présentés ci-après.

1.2.13.1 Principe inventif

Altshuller a dégagé à partir des brevets étudiés 40 principes d'innovation (cf. tableau 1.3). L'annexe A fournit des informations plus complètes sur ces principes.

N.	Principe	N.	Principe
1	Segmentation	2	Extraction
3	Qualité locale	4	Asymétrie
5	Combinaison	6	Universalité
7	Placement intérieur « poupées russes »	8	Contrepoids
9	Action inverse préliminaire	10	Action préliminaire
11	Compensation ou protection préliminaire	12	Equipotentialité
13	Inversion	14	Sphéricité
15	Mobilité	16	Action partielle ou excessive
17	Changement de dimension	18	Vibration mécanique
19	Action périodique	20	Continuité d'une action utile
21	Grande vitesse	22	Application bénéfique d'un effet néfaste
23	Asservissement	24	Intermédiaire
25	Self-service	26	Copie
27	Ephémère et bon marché	28	Remplacer les Systèmes mécaniques
29	Systèmes pneumatiques et hydrauliques	30	30 Membrane flexible et film mince
31	Matériau poreux	32	Changement de couleur
33	Homogénéité	34	Eliminer récupérer
35	Changement de paramètre	36	Changement de phase
37	Dilatation thermique	38	38 Oxydants puissants
39	Environnement inerte	40	Matériaux composites

Tableau 1.3: Principes inventifs

Pour chaque couple de paramètres (cf. § 1.2.12) entrant en conflit, Altshuller a défini les principes inventifs recommandés pour résoudre ce conflit. Ces recommandations sont présentées dans une matrice appelée la matrice de contradiction (cf. tableau 1.4). Ses lignes représentent les paramètres à améliorer et ses colonnes représentent les paramètres indésirables. A l'intersection de chaque ligne et colonne se trouvent les principes inventifs recommandés. Le tableau 1.4 présente un extrait de cette matrice. La matrice complète est donnée dans [23].

Les paramètres indésirables / Les paramètres à améliorer		1 poids d'un objet mobile	2 poids d'un objet fixe	…	10 Stabilité	…	38 Degré d'automatisation	39 Productivité
7	**Volume d'un objet mobile**	2, 26, 29, 40	-		15, 35, 36, 37		35, 34, 16, 24	10, 6, 2, 34
8	**Forme**	8, 10, 29, 40	15, 10, 26, 3		33, 1, 18, 4		15, 1, 32	17, 26, 34, 10
⋮								
39	**Productivité**	35, 26, 24, 37	28, 27, 15, 3,		28, 15, 10, 36		5, 12, 35, 26	-

Tableau 1.4: Extrait de la matrice de contradiction

1.2.13.2 Principe de séparation

Un principe de séparation permet de résoudre une contradiction physique (cf. § 1.2.11.2). Les principes essentiels sont les quatre principes suivants:

1. Séparation dans l'espace : le système est partitionné en sous-systèmes. Chacun effectue une des fonctions contradictoires.

2. Séparation dans le temps : le système est planifié dans le temps pour réaliser chaque fonction en conflit dans un temps spécifique.

3. Séparation en (partie/tout) systèmes : par cette approche, des fonctions contradictoires peuvent être réalisées par des systèmes séparés. Une fonction contradictoire avec la fonctionnalité d'un système peut être réalisé par :

   - un super-système ;
   - un sous-système ;
   - un système alternatif (différent) ;

- un système contradictoire.

4. Séparation des conditions : ce principe résout des contradictions où une fonction utile est réalisée quand une condition spéciale existe. La résolution se fait par la modification du système ou de son environnement.

Des exemples d'application de ces principes sont présentés dans [19, 7].

1.2.13.3 Effet scientifique

Un effet représente une relation cause-conséquence (ou entrée-sortie [22]). La cause (l'entrée) et la conséquence (la sortie) sont des fonctions. La fonction de sortie d'un effet doit être prise en compte dans le développement du système pour le simplifier et pour ne pas nuire à son fonctionnement [7].

Un effet scientifique est un effet observé et formulé par une loi mathématique, physique, chimique ou géométrique. Un exemple d'un effet physique est la loi de Coulomb décrite comme suit : « Deux objets électriquement chargés par des charges similaires créent une force de répulsion. Deux objets électriquement chargés par des charges différentes créent une force d'attraction ». Les bases de connaissances des outils d'innovation commercialisés (cf. § 1.3) présentent des milliers d'effets scientifiques [24, 25]. Un extrait de ces effets de la base de TechOptimizer (cf. § 1.3.1) est présenté dans l'annexe B.

1.2.13.4 Solution innovante générique (*standard solution*)

La méthode TRIZ présente 76 solutions généralisées utilisées pour résoudre des problèmes d'innovation. Ces solutions génériques sont représentées comme des recommandations permettant de compléter ou d'améliorer un système. Elles sont classifiées aussi dans cinq classes principales [6] :

1. Composition et décomposition des modèles champ-substance ;
2. Amélioration et évolution des modèles champ-substance ;
3. Transition vers un super système ou vers le micro niveau ;
4. Mesure et détection ;

5. Aides.

Nous donnons ci-après deux exemples de solutions.

- *Solution 1-1-6* : si un effet minimum (optimum, mesurable,…) d'une action est requis, mais qu'il est difficile ou impossible à obtenir dans les conditions du problème, utiliser une action maximum tout en enlevant la partie excessive de l'action ; un excès de substance est enlevé par un champ, alors qu'un excès de champ est enlevé par une substance.

- *Solution 5-4-2* : s'il est nécessaire d'obtenir un effet fort en sortie du système, alors que l'effet d'entrée est faible, la substance transformante est placée dans des conditions proches d'un état critique ; l'énergie est stockée dans la substance et le signal d'entrée agit comme un déclencheur.

1.3 Outils informatiques de résolution des problèmes d'innovation

Plusieurs logiciels commerciaux basés sur la méthode TRIZ sont disponibles pour assister l'utilisateur dans l'analyse et la résolution de son problème. Dans cette section nous présentons les plus connus.

1.3.1 Goldfire innovator

C'est une nouvelle version de TechOptimizer [25] développée par *Invention Machine Corporation* [18, 22]. Il comporte trois composants essentiels :

1.3.1.1 Optimiseur (*Optimizer*)

Ce module assiste l'utilisateur pour optimiser la conception d'un système technique. Il est basé sur trois types de *workflows* paramétrés par l'utilisateur :

- Le premier *workflow*, « *Improve Existing System* », permet d'améliorer un système existant par une analyse de la cause première des problèmes (« *Root Cause Analysis* »). Ainsi, cet outil permet de se focaliser sur le bon problème à résoudre. La formalisation de ce problème est basée sur un modèle appelé cause-effet dans lequel

les événements indésirables sont définis par l'utilisateur. Cet outil aide aussi à choisir la meilleure conception à partir de plusieurs variations données.

- Le deuxième *workflow*, « *New Device Analysis* », est un outil d'analyse de nouveaux appareils. Ce *workflow* permet de créer le modèle de fonctionnement de l'appareil et d'identifier les solutions possibles pour réaliser les fonctions exigées. Cet outil aide aussi à sélectionner la meilleure configuration de la conception à partir de plusieurs simplifications et paramètres donnés par l'utilisateur.

- Le troisième *workflow*, « *System Benchmarking* », aide à comparer systématiquement plusieurs systèmes et à réunir leurs meilleures caractéristiques dans un système hybride optimisé.

1.3.1.2 Module de recherche (*researcher*)

Ce module permet de rechercher sémantiquement des ressources de connaissances. Son interface supporte la recherche par des requêtes composées en langage naturel (anglais ou français cf. figure 1.4) ou en langage logique (cf. figure 1.5).

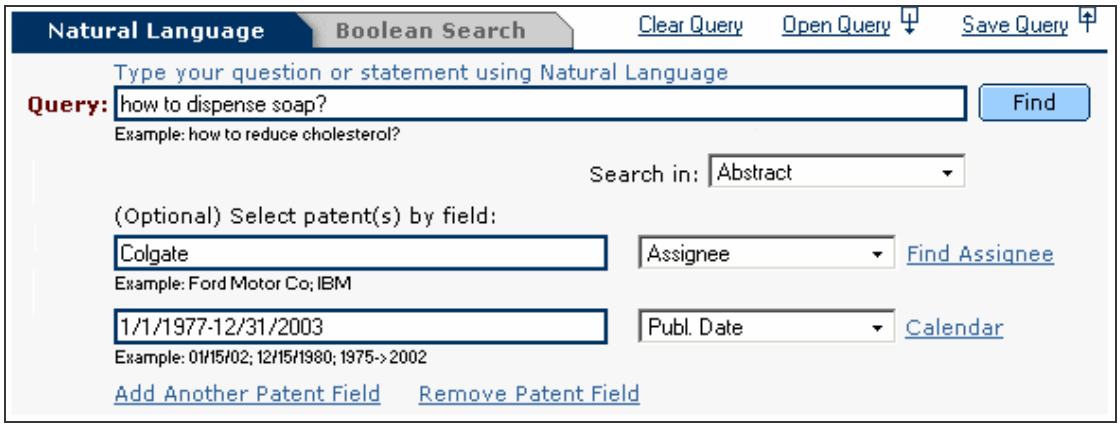


Figure 1.4: Interface des requêtes en langage naturel

Les ressources de connaissances peuvent être repérées :

1. Dans le Web ;

2. Dans une base de documents personnels ;

3. Dans les collections de brevets comme USPTO, EPO, WIPO et JPO. Ces collections sont pré-indexées, résumées et hebdomadairement mises à jour par Invention Machine ;

4. Dans une base de 8.000 effets scientifiques et exemples en anglais.

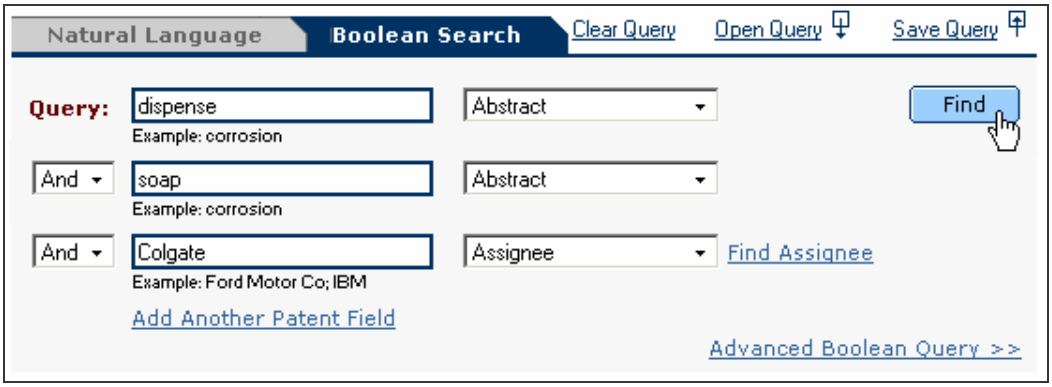


Figure 1.5: Interface des requêtes en langage logique

1.3.1.3 Analyse de la tendance de l'innovation (*Innovation Trend Analysis*)

Ce module permet à l'utilisateur d'extraire des connaissances sur des situations d'entreprises à partir des collections des brevets indexés et résumés par Invention Machine. Les connaissances extraites sont :

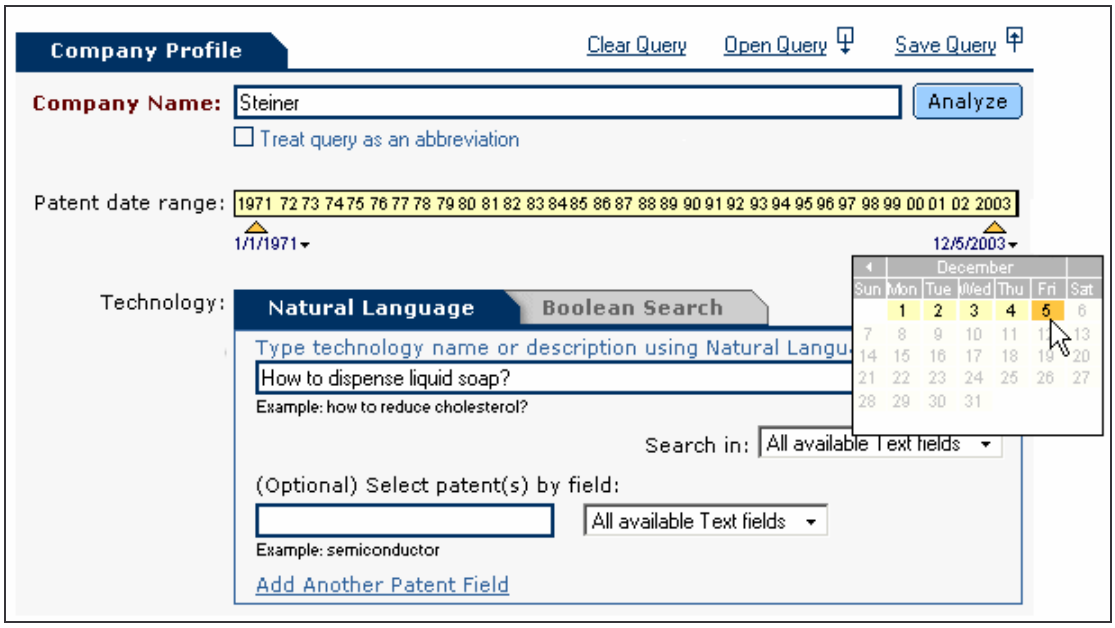


Figure 1.6: Interface d'analyse de la tendance de l'innovation

1. Le profil d'une compagnie : il permet de reconnaître la tendance inventive et technologique d'une compagnie (cf. figure 1.6).

2. L'analyse compétitive: elle permet de comparer le profil inventif de plusieurs compagnies (jusqu'à cinq).

3. L'analyse technologique : elle permet de reconnaître la tendance des technologies apparues.

4. La citation des brevets : elle permet de reconnaître le propriétaire de chaque technologie apparue.

1.3.1.4 Résolution des contradictions techniques

Goldfire Innovator fournit une interface spécifique pour résoudre les contradictions techniques et physiques. Les principes inventifs permettant de résoudre une contradiction technique sont repérables en sélectionnant les paramètres en conflit (cf. figure 1.7). Le système filtre et propose à l'utilisateur jusqu'à quatre recommandations illustrées par des exemples.

1.3.1.5 Résolution des contradictions physiques

La base de connaissance de Goldfire Innovator intègre cinq principes dits « de haut niveau ». Ces principes sont :

1. Séparation dans le temps (*Separation in time*) ;

2. Séparation dans l'espace (*Separation in space*) ;

3. Séparation dans l'espace des phases (*Separation in phase space*) : ce principe affirme qu'un paramètre *P* peut avoir différentes valeurs par rapport à deux espaces des phases. Ce principe peut être vu comme une généralisation des autres principes de séparation représentés dans un espace des phases ;

4. Séparation en niveaux système/sous-système (*Separation of system-subsystem levels*) ;

5. Séparation en niveaux système/super-système (*Separation of system-sypersystem levels*).

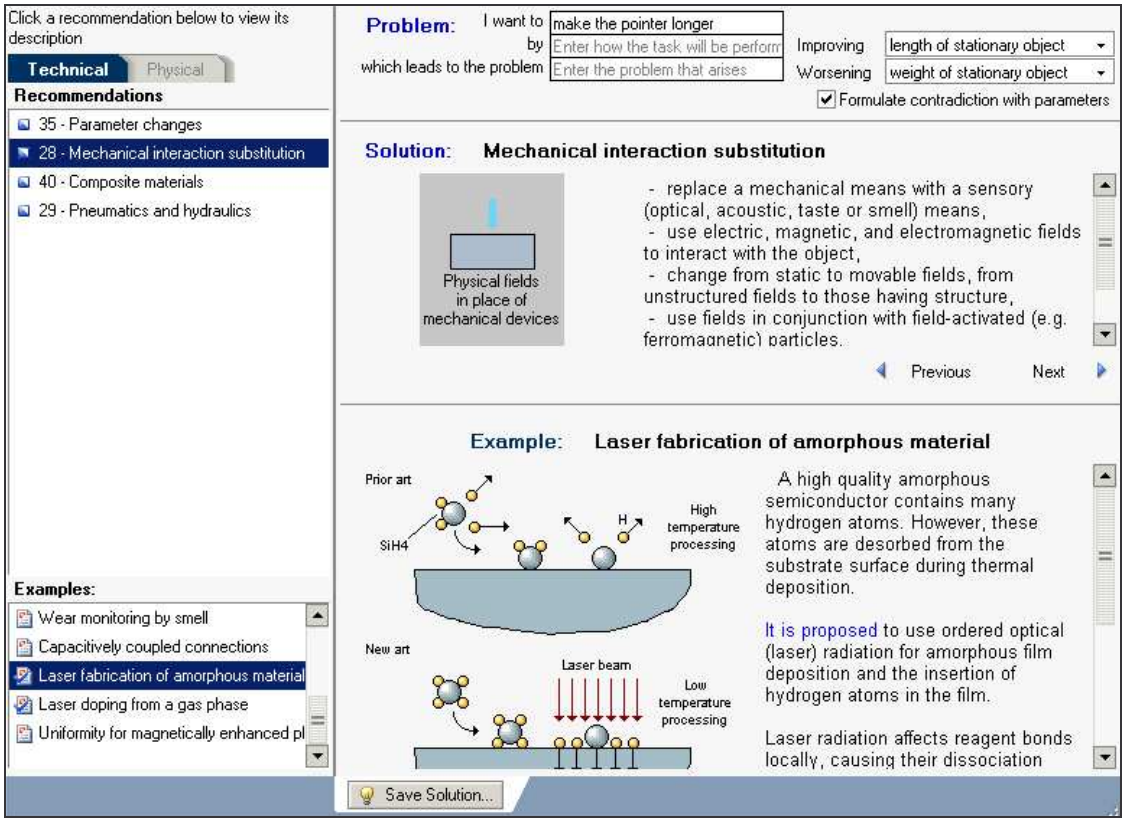


Figure 1.7: Interface des principes

Dans la même interface des principes (cf. figure 1.7), les principes de séparation et leurs exemples sont repérables exactement comme les principes inventifs.

1.3.1.6 Effets scientifiques

La base de Goldfire Innovator permet l'accès à 8000 théorèmes, lois et processus formulés dans les disciplines scientifiques. Ces effets sont organisés par fonction de sortie, par groupe de fonctions, par ressource et par industrie (Automobile, Microélectronique, Lasers & Optique). Ils sont associés à des exemples illustratifs. La figure 1.8 présente l'interface spécifiée pour l'exploration de ces effets et de leurs exemples. Un extrait de ces effets est listé dans l'annexe B.

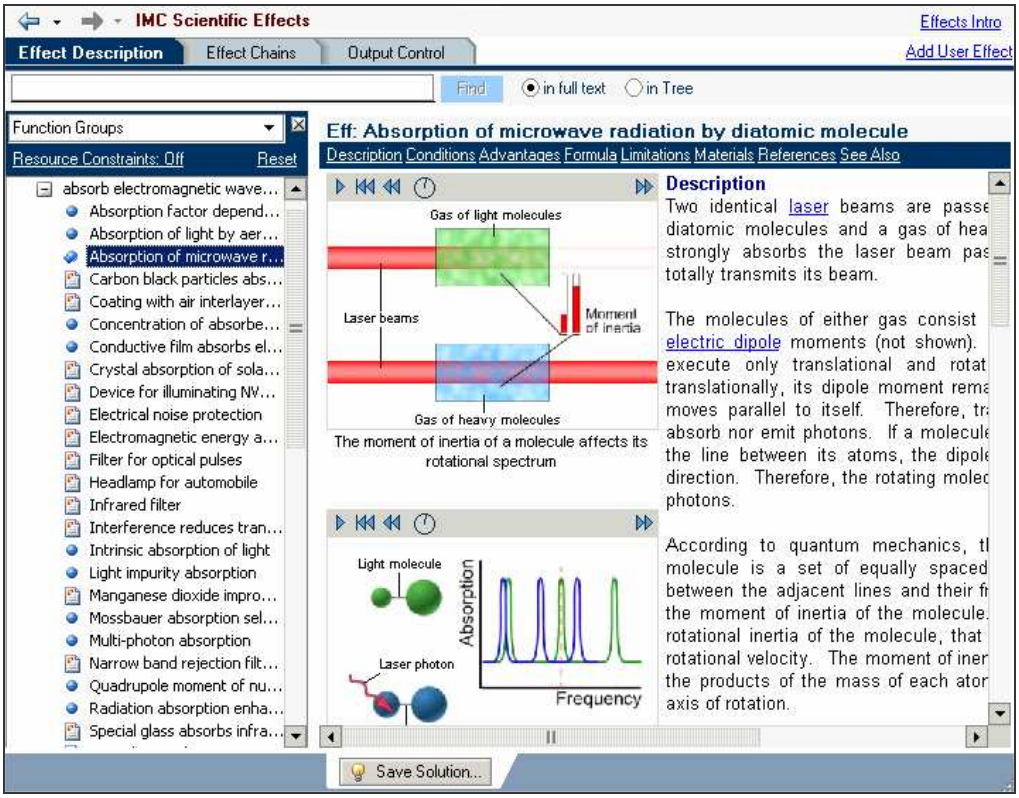


Figure 1.8: Interface des effets scientifiques

1.3.2 Innovation WorkBench

Le logiciel IWB [19] d'Idéation International comporte sept modules :

1. Innovation Situation Questionnaire (ISQ) ;
2. Problem Formulator ;
3. Navigator ;
4. System of Operators ;
5. Results Analysis ;
6. Innovative Illustrations Library ;
7. Innovation Guide.

1.3.3 TRIZ Explorer

La base de connaissances de TRIZ Explorer [20] contient 5 sections :

1. Les principes d'innovations ;

2. Les solutions innovantes génériques (Inventive Standards) ;

3. Les effets scientifiques ;

4. Les ressources de TRIZ sur l'Internet ;

5. Les bases de connaissances privées (ou utilisateur).

L'utilisateur peut étendre toutes les sections de la base par des nouveaux concepts et ressources.

1.3.4 Méthode d'aide à l'innovation MAL'IN

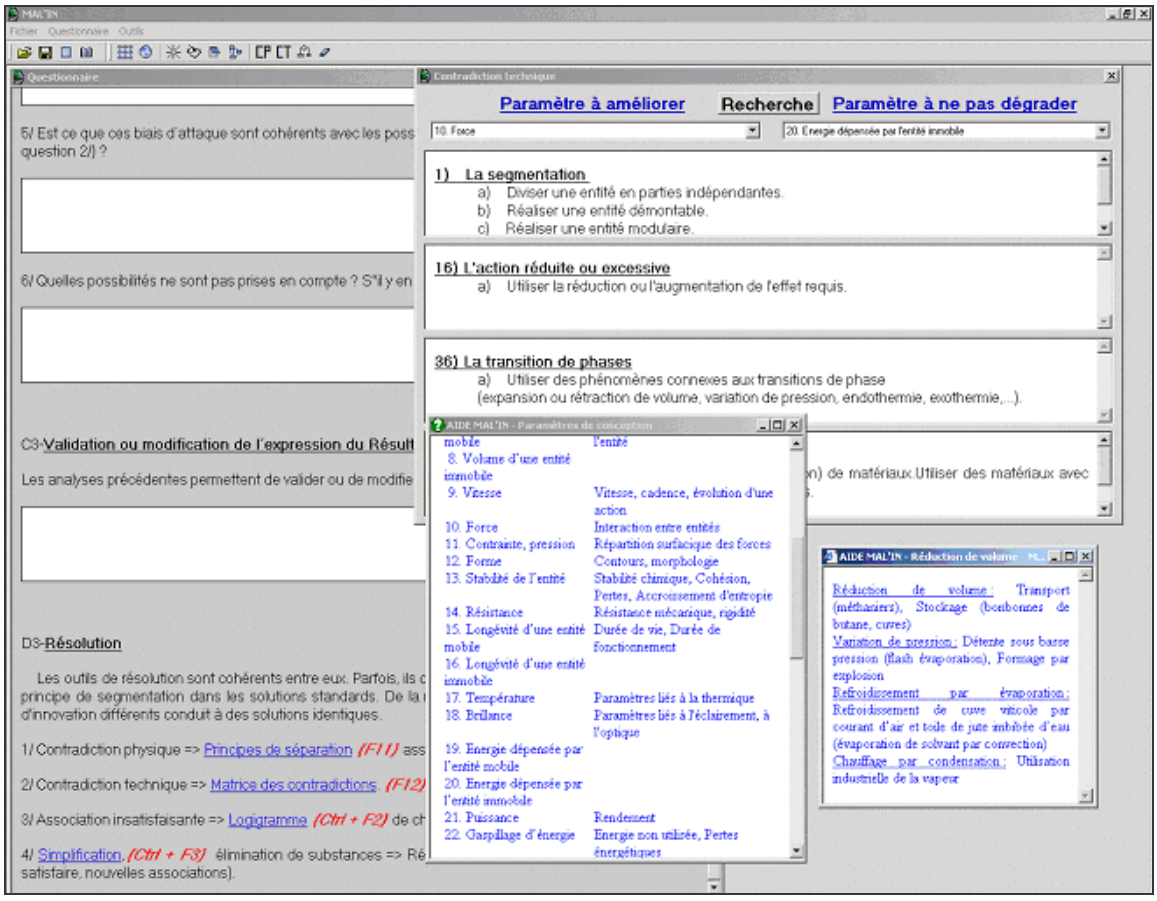


Figure 1.9: Interface des effets scientifiques

MAL'IN (cf. figure 1.9 et [88]) est un outil devéloppé à l'université Bordeaux I. Par cette outil, les développeurs essayent d'adapter la méthode TRIZ pour un usage par des non-spécialistes, en se basant essentiellement sur :

1. Un vocabulaire et des outils de l'analyse fonctionnelle ;

2. Une réécriture des outils TRIZ pour focaliser la recherche d'idées ;

3. Des exemples ciblés pour les bureaux d'études ;

4. Des outils d'analyse de graphes.

1.3.5 Autres outils

Il existe aussi d'autres outils d'aide à l'innovation comme CreaTRIZ [26], TRISolver [27]. Basé sur la méthode TRIZ, chaque outil représente et implémente une nouvelle base de connaissances et offre un certain nombre de services pour assister l'utilisateur dans la conception des systèmes et la résolution des problèmes apparus. Ainsi la compétition principale entre ces outils est de faciliter la représentation et la formulation des problèmes d'innovation ainsi que le repérage sémantique d'une solution à chaque problème dans la base de connaissances ou dans une base documentaire (les bases de brevets, le Web, la base locale d'une entreprise).

1.4 Problématique étudiée

Les outils d'aide à l'innovation présentés dans la section précédente ne sont pas suffisamment élaborés pour aider à la résolution automatique/semi-automatique des problèmes. En phase de conception, l'utilisateur doit lui-même identifier et formuler le problème d'innovation et rechercher des solutions. L'aide à la recherche de solutions menée par ces outils est limitée au contenu de la base de connaissances. Le contenu de la base doit être mis à jour périodiquement par la compagnie productrice.

Or de très nombreuses informations et connaissances sont disponibles sur le Web. Il devient donc intéressant et même indispensable d'envisager des approches afin d'automatiser l'extraction de connaissances à partir du Web et l'enrichissement des bases de connaissances relatives à l'innovation. L'ambition est d'évoluer vers des outils plus « intelligents » dans l'aide à la résolution des problèmes capables d'exploiter la sémantique d'un domaine pour la recherche de solutions.

L'innovation comme présentée au début de ce chapitre est une activité qui intéresse plusieurs domaines de connaissances. Ainsi, des opérateurs d'innovation peuvent être définis dans ces domaines et intégrés dans une base de connaissances. Dans l'état actuel des outils d'aide à l'innovation, l'expert peut mémoriser de nouveaux opérateurs dans une base de connaissances et leur associer manuellement des exemples. Ensuite, ces opérateurs seront accessibles par l'expert et son groupe de travail. Notre ambition est de concevoir une représentation sémantique globale de ces opérateurs d'innovation par une ontologie d'innovation. L'expert peut alors se baser sur cette ontologie pour définir les nouveaux opérateurs. Des logiciels d'aide à l'innovation peuvent accéder à ces opérateurs pour automatiser les deux tâches suivantes :

1. Association des exemples pertinents à ces opérateurs à partir des textes et des brevets publiés sur le Web ou dans une collection de documents.

2. Utilisation de ces opérateurs et des exemples associés pour aider à résoudre les problèmes d'innovation.

Dans cette thèse, nous avons conçu un système d'aide à l'innovation en vue d'automatiser ces deux tâches. Plus précisément, nous nous sommes attachés à concevoir une approche pour associer des exemples à des opérateurs d'innovation définis dans une base de connaissances d'innovation.

Chapitre 2

Etat de l'Art

Dans ce chapitre, nous présentons l'extraction d'information (EI) (cf. § 2.1) et les systèmes question-réponse (SQR) (cf. § 2.2). Ces deux types de systèmes nous sont apparus comme étant les plus proches de nos préoccupations. L'extraction d'information est une technique complémentaire de la recherche documentaire qui a pour objectif d'extraire des informations spécifiques d'une collection de textes en langue naturelle (LN). Un SQR a pour objectif d'extraire à partir de textes en langue naturel LN des réponses à des questions posées par l'utilisateur, et non pas une liste de documents susceptibles de contenir cette réponse.

2.1 Extraction d'information (EI)

2.1.1 Définition et positionnement

L'extraction d'information (EI) est l'activité qui consiste à rechercher automatiquement des informations dans des textes en langue naturelle LN et à représenter ces informations (cf. figure 2.1.*a*) sous la forme de motifs structurés conformément à une banque de données (cf. figure 2.1.*b*) [28, 29, 65, 85]. Les motifs à extraire sont souvent définis par un ou plusieurs formulaires (cf. § 2.1.4). Ces motifs concernent souvent des entités nommées EN (en gras dans la figure 2.1.*a*) désignant des personnes, des organisations et des lieux, ainsi que des marques, des dates et des unités numériques. Ainsi, l'EI consiste à remplir une instance de ce formulaire à partir des textes en langue naturelle LN. Ce remplissage nécessite le repérage des EN, l'annotation de leur type et leur mise en relation avec les motifs du formulaire. Un système d'EI (cf. figure 2.2) sert par conséquent à repérer dans les textes en langue naturelle LN des EN décrites dans le formulaire et à alimenter leurs instances dans une base spécifique. A titre d'exemple, des systèmes d'EI peuvent être conçus pour surveiller les actualités dans un cadre de veille concurrentielle ou scientifique et technique [28].

L'EI est différente de la recherche d'information RI (Information Retrieval) qui vise à repérer dans une collection de documents (textes, fichiers multimédias, bases de données,…) ceux qui correspondent le mieux à une requête donnée [64]. Mais, les deux techniques sont complémentaires. La RI est nécessaire dans l'extraction pour repérer les documents pertinents à l'application [54].

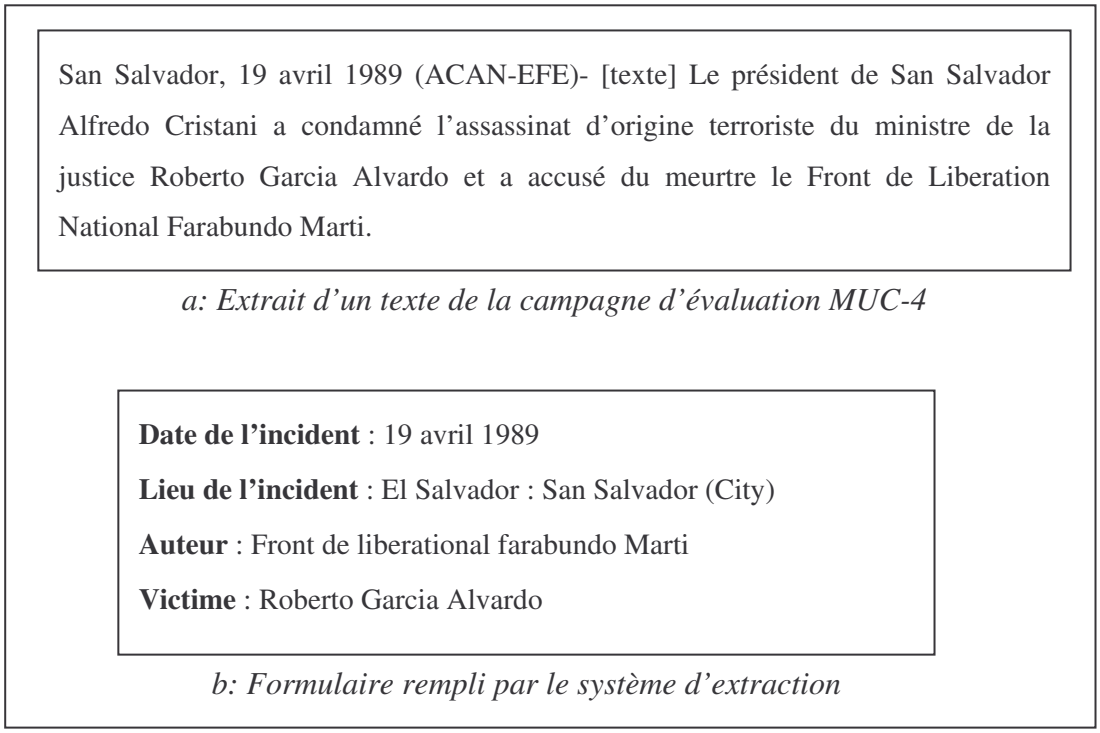


Figure 2.1: Exemple représentatif de résultats d'un système d'extraction d'information

L'EI peut être considérée comme une forme simplifiée de la compréhension de textes en langue naturelle [29]. Ses approches essaient de réduire et de simplifier le traitement linguistique sur les textes, en l'appliquant d'une manière superficielle et sélective sur les textes. Les passages candidats sont analysés par une grammaire réduite (cf. § 2.1.3.1) dépendant de la tâche d'extraction (cf. § 2.1.4). Seule une partie relativement minime du texte peut nécessiter une analyse approfondie : 10 à 20 % en moyenne avec une diminution à 3% pour certains textes techniques [29].

Dans le traitement sémantique, le sens d'un terme est défini à partir d'un ensemble fixé d'objets sémantiques : des entités, des relations entre entités et des événements réalisés par les entités. Ces objets sont définis par un ensemble de types ayant tous une structure régulière dans un domaine d'application [91].

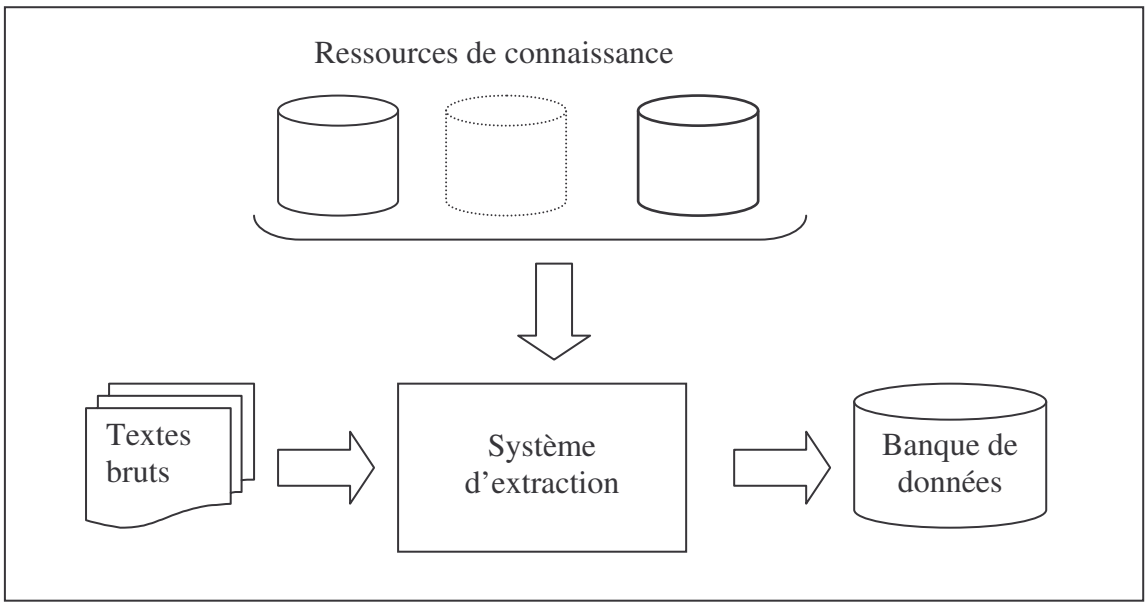


Figure 2.2: Vue générale d'un système d'extraction d'information

Par conséquent, l'EI a des limites et des difficultés. Cette technique ne peut pas garantir la validité de l'information extraite. Le traitement superficiel et local des textes focalisant sur certains éléments néglige volontairement ou involontairement beaucoup d'éléments qui nécessiteraient des traitements complémentaires complexes pour réduire les ambiguïtés [29]. Les indicateurs d'extraction sont sujets à des variations et une simple variation dans un texte peut perturber l'extraction et introduire des bruits. Par exemple, entre *venture between* et *ventures between*, il y a une grande différence de signification. La première forme est utilisée pour détecter des accords signés entre entreprises et l'autre fait référence au verbe *to venture* [28,34]. En revenant à la figure 2.1, on remarque aussi que l'extraction entraîne :

- La perte de la nuance linguistique de la relation ou de l'événement [28]. L'accusé dans le texte de la figure 2.1.*a* est devenu criminel dans le formulaire extrait de la figure 2.1.*b*.

- L'omission des informations essentielles [28]. Dans le texte cette information est l'accusation dont l'acteur est le présidant du San Salvador. Par contre, cette information n'apparaît pas du tout dans le formulaire extrait.

Plusieurs applications de l'EI et de la fouille de texte [28] pourraient être automatisées par la résolution de ces difficultés et limites.

Actuellement, les applications de l'extraction d'information sont nombreuses et variées. On peut citer : le suivi des analyses des patients [31], le suivi de fusions d'entreprises [86], le suivi de la localisation subcellulaire d'une protéine [87]. De nombreuses autres applications sont citées dans les campagnes d'évaluation MUC des systèmes d'extraction (cf. § 2.1.5) dans les domaines du terrorisme, de la micro-électronique, du changement de direction d'entreprises.

Les applications de la fouille de textes les plus concernées par l'extraction d'information sont l'acquisition automatique des ressources de connaissances (dictionnaires, ontologies, méta données, bases de règles) (cf. § 2.1.3) et le développement des systèmes de question-réponse (cf. § 2.2).

2.1.2 Approches générales de l'extraction d'information

L'élément clé de l'EI est l'élaboration des règles d'extraction permettant de repérer dans les textes en langue naturelle des éléments sémantiquement définis dans un domaine d'application [61, 91]. Des patrons linguistiques sont utilisés pour mettre en relation syntaxique des classes de termes définis dans des ressources de connaissances (lexiques dictionnaires, ontologies, thésaurus). La mise en correspondance des patrons avec un texte (cf. figure 2.3) nécessite souvent une analyse linguistique du texte (cf. § 2.1.3.1), un repérage des entités nommées (EN) (cf. § 2.1.4.1) et une résolution de certains types de co-références (cf. § 2.1.4.2). Ensuite les règles d'extraction sont utilisées pour extraire les éléments pertinents et les remplir dans le formulaire.

La mise au point des règles d'extraction et des autres ressources de connaissances nécessaires pour l'extraction peut être effectuée manuellement par des experts (linguistes et ingénieurs de connaissances), en semi-automatique ou en automatique en utilisant des systèmes d'apprentissage (acquisition automatique cf. § 2.1.3.2).

- Les approches manuelles comme pour le système FASTUS [49,68] sont « taillées sur mesure ». Elles élaborent des ressources de connaissances extrêmement spécifiques à leur domaine d'application. Elles demandent beaucoup de compétence, de temps et d'efforts et sont très difficiles à mettre au point [29, 85]. Leurs performances dépendent du corpus et du domaine d'application. Elles sont perturbées par le style, la syntaxe et/ou la sémantique des corpus [28]. Leur adaptation à un nouveau domaine ou corpus par des personnes très

compétentes peut demander des mois d'effort pour mettre à jour les règles d'extraction et les ressources de connaissances [54]. Des systèmes comme Pinocchio [55] disposent d'un environnement de modules réutilisables facilitant ces mises à jour. Pour ce même objectif, on note de nouvelles approches [60] qui construisent leurs ressources de connaissances en se basant sur des ontologies, des techniques et des outils développés pour le Web sémantique.

➢ Les approches fondées sur l'apprentissage sont développées pour faciliter l'adaptation des systèmes à de nouvelles applications. L'acquisition de ressources de connaissances dans ce type d'approches se base essentiellement sur un corpus d'entraînement. L'entraînement construit automatiquement un modèle d'extraction prenant la forme d'un dictionnaire de patrons symboliques d'extraction (AutoSlog [50], CRISTAL [31] et PALKA [51] et WHISK [61], Yangarber *et al.* [91]), de règles logiques [32, 63, 94] ou d'un modèle numérique comme le modèle de Markov caché (*Hidden Marcov Model*) [33, 87].

Classiquement, l'acquisition se fait à partir d'un corpus annoté (AutoSlog [50], CRISTAL [31], SRV [63], $(LP)^2$ [32]). Ces systèmes ont l'inconvénient qu'ils ne sont pas toujours portables à une nouvelle application. Un corpus d'entraînement d'une bonne qualité et d'une taille suffisante pour l'apprentissage est rarement disponible. Son élaboration est complexe et demande des mois de travail de plusieurs analystes humains avec le risque de subjectivité inhérente à l'annotation manuelle [28]. L'annotation manuelle des textes est une des origines de la complexité. Apprendre à choisir et délimiter les termes à annoter demande beaucoup d'expérience [30]. Prenons à titre d'exemple l'annotation des EN dans une application. Un expert du domaine qui n'est pas le développeur du système doit les repérer dans les phrases nominales du corpus. Mais, que doit-il annoter exactement dans ces phrases? Est-ce qu'il doit annoter les modifieurs[1]? Les articles? Comment annoter ces entités dans les conjonctions?

D'autres approches ont été proposées pour surmonter cette complexité d'annotation. A titre d'exemple, le système AutoSlog-TS [30] crée un dictionnaire des patrons d'extraction en se basant sur deux collections de textes utilisées comme corpus d'entraînement : une collection pour les textes pertinents et l'autre pour les textes non pertinents. Les systèmes RAPIER [58, 85, 94], WHISK [61], PALKA [51], LIEP [86] utilisent des méthodes

[1] Un modifieur est un nom qui modifie un autre dans un groupe nominal. En français, il vient en général après la proposition 'de' comme 'information' dans '*extraction d'information'*. En anglais, il vient après 'of' ou directement avant un nom principal comme 'information' dans l'expression 'Information extraction'

d'apprentissage supervisé. Par ce type d'apprentissage, le système participe à la sélection et/ou à l'annotation des textes à partir d'un corpus non annoté. Cette approche permet donc de réduire le nombre des exemples à annoter manuellement pour arriver à un niveau défini de performance. D'autres systèmes apprennent automatiquement les patrons d'extraction à partir d'un corpus non annoté, en se basant sur une souche des mots (*seed words*) (MetaBoostrapping [93]) ou une souche des patrons (ExDisco [92]) définie à l'amorçage du système.

Cependant, l'adaptation de ces approches à un nouveau domaine d'application pose aussi des problèmes. Les systèmes existants sont des projets de recherche en laboratoires. Ces systèmes restent expérimentaux et n'ont pas été testés à grande échelle. L'acquisition automatique des règles d'extraction rencontre des limites sérieuses. Certains domaines nécessitent des connaissances que les systèmes d'extraction ne peuvent pas acquérir. Par exemple, l'acquisition des classes de termes dans des hiérarchies profondes est encore très difficile (cf. § 2.1.3.2).

2.1.3 Etapes préparatoires de l'extraction d'information

Dans un système d'extraction d'information (cf. figure 2.3), on peut distinguer en général trois étapes principales: le repérage des EN, la mise en relation des entités et le remplissage du formulaire [28]. Les tâches effectuées dans ces étapes sont présentées dans la section suivante. Dans cette section, on présente les étapes préparatoires pour l'extraction comme l'analyse linguistique et l'acquisition de ressources de connaissances.

2.1.3.1 Analyse linguistiques de textes

L'analyse linguistique est une étape préparatoire nécessaire pour toute tâche d'EI dans des textes en langue naturelle (textes libres, semi-structurés et structurés [61]). Elle est aussi nécessaire pour l'acquisition automatique des ressources de connaissances (cf. § 2.1.3.2).

Dans l'extraction, cette analyse permet de résoudre certaines ambiguïtés possibles dans l'utilisation des ressources de connaissances (cf. § 0). Cette analyse peut être morphologique et lexicale permettant de mettre en évidence la catégorie (POS) du mot (n, pn, v, adv, adj, pré…) et sa forme canonique (lemme), ce qui facilite par la suite le repérage des informations complémentaires dans les dictionnaires, les réseaux sémantiques ou les ontologies. Elle peut être syntaxique permettant aussi d'identifier des fonctions grammaticales des termes ou des

classes de termes dans les phrases (sujet, verbe, objet direct, objet indirect, modifieur). Cette analyse syntaxique est très utile pour l'extraction des relations binaires et des événements (cf. § 2.1.4.4 et § 2.1.4.5).

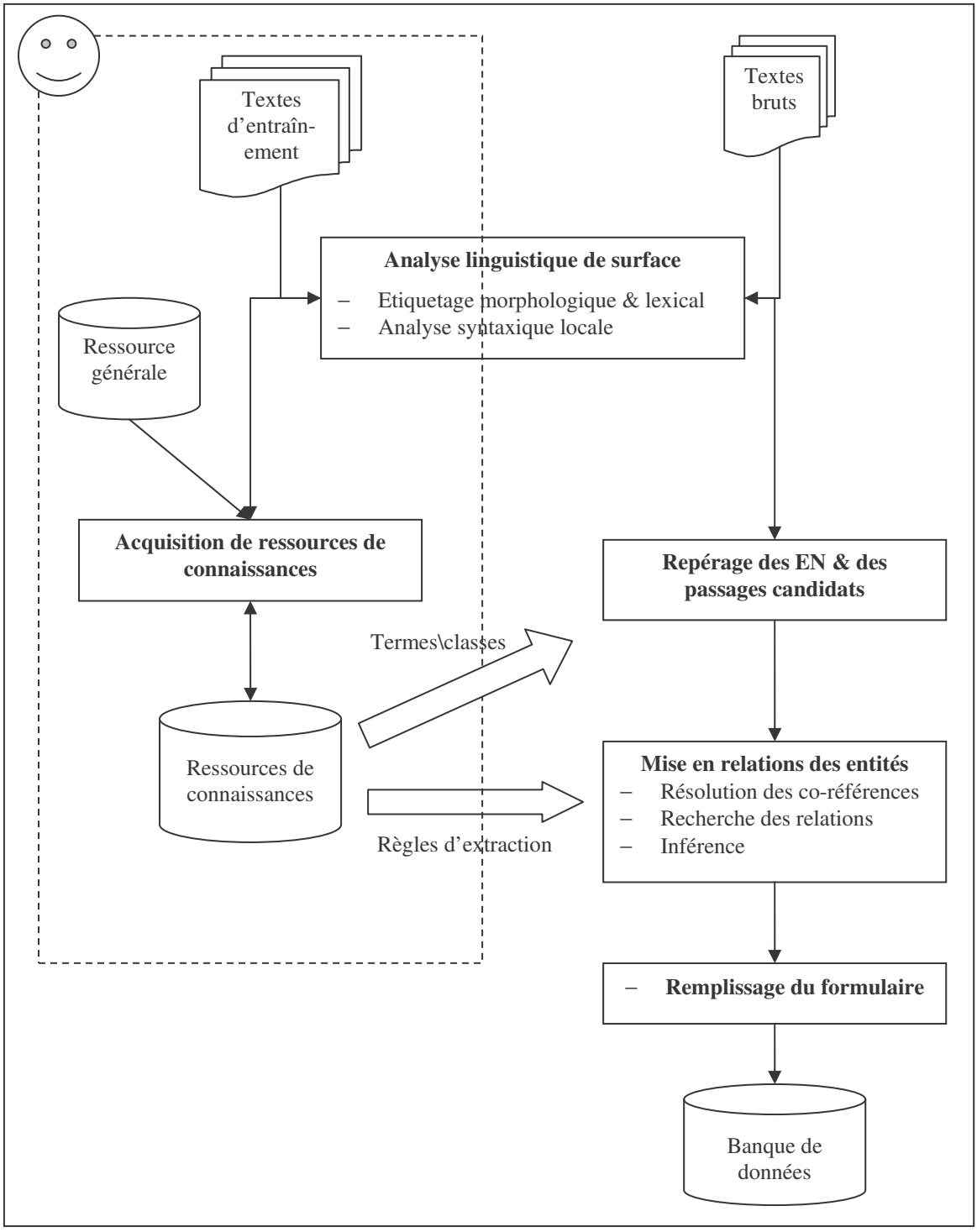


Figure 2.3: Etapes de l'extraction d'information

Le choix de la profondeur de l'analyse syntaxique constitue un problème dans toutes les approches d'extraction et d'acquisition [54, 85]. Une analyse syntaxique complète (cf. figure 2.4) n'est pas conseillée parce qu'elle est difficile à faire localement, car la grammaire ne couvre généralement pas tous les cas de figures. En plus, les détails fournis par cette analyse compliquent l'acquisition des règles et l'extraction d'information. Beaucoup d'entre eux sont négligés pour faciliter l'implémentation et l'exécution des tâches d'extraction (cf. § 2.1.4). Par ailleurs, une analyse trop superficielle entraîne un risque d'inadaptation des règles pour certains domaines techniques.

```
- SUBJ_PASSIVE(quantifié,il)
- OBJ(rajeunir,l'âge)
- OBJ(quantifié,risque)            //incorrecte
- DEEPOBJ(quantifié,il)
- VMOD(quantifié,correctement)
- VMOD(rajeunir,considérablement)
- NMOD_POSIT1(l'âge,surface)
- NMOD_POSIT1(surface,Mars)
- NMOD(risque,rajeunir)            //incorrecte
- PREPOBJ(Mars,de)
- PREPOBJ(surface,de)
- PREPOBJ(rajeunir,de)
- DETERM_DEF(surface,la)
- DETERM_DEM(résultat,Ce)
- AUXIL_PASSIVE(quantifié,sera)
- CONNECT(quantifié,quand)
- 0>GROUPE{NP{Ce résultat}, SC{BG{quand} NP{il} FV{sera
  correctement    quantifié}},    NP{risque}    IV{de
  considérablement  rajeunir}  NP{l'âge}  PP{de  NP{la
  surface}} PP{de NP{Mars}} .}
```

Figure 2.4: Analyse syntaxique de la phrase « Ce résultat, quand il sera correctement quantifié, risque de considérablement rajeunir l'âge de la surface de Mars.» par XIP Parser [106]

Des systèmes d'extraction font une analyse syntaxique superficielle par le moyen de procédures d'analyses basées sur une grammaire réduite. Les patrons présentés dans la figure 2.5, par exemple, sont utilisés par FASTUS [49,68] pour associer les sujets à leurs verbes. Le premeir patron est utilisé lorsque le sujet est suivi par une préposition et l'autre lorsqu'il est suivi par une clause. AutoSlog [50] et AutoSlog-TS [30] emploient des patrons linguistiques

heuristiques pour acquérir les règles d'extraction à partir du corpus d'entraînement ainsi que pour l'extraction d'information. WHISK [61], CRISTAL [31] effectuent cette analyse en utilisant l'analyseur syntaxique BADGER [104]. RAPIR [85] se base seulement sur l'analyse lexicale pour l'acquisition de ses règles d'extraction.

Pour l'analyse morphologique et lexicale, des outils spécifiques (encore appelés *POS taggers*) comme TreeTagger [105], Xerox POS tagger [106] et INTEX [75] peuvent être utilisés pour l'anglais, le français et d'autres langues.

```
Subject {Preposition NounGroup}*
    VerbGroup
Subject Relpro {NounGroup | Other}*
    VerbGroup {NounGroup | Other}*
    VerbGroup
```

Figure 2.5: Extrait de la grammaire utilisée dans le système FASTUS

2.1.3.2 Acquisition de ressources de connaissances

Dans l'EI, deux types de connaissances sont nécessaires pour l'extraction. Ce sont des classes sémantiques de termes qui définissent les arguments de l'information visée et des règles d'extraction qui définissent leur contexte [29, 61]. Les classes de termes (cf. figure 2.3) sont utilisées pour faciliter l'acquisition des règles d'extraction ainsi que le repérage des passages candidats dans les textes. Les règles d'extraction représentent des contraintes syntaxiques (grammaticales) sur les classes. Elles prennent souvent la forme de patrons linguistiques.

Comme présentée dans la section 2.1.2, la construction manuelle de ces ressources est très difficile. L'ambition actuelle est de faciliter cette tâche. Plusieurs approches et outils d'acquisition sont suggérés pour automatiser l'acquisition de ces ressources.

Acquisition de classes de termes

L'acquisition de classes de termes peut être faite avec l'acquisition des règles ou au cours d'une étape préparatoire. Actuellement, un nombre de plus en plus grand d'approches sont intéressées par cette acquisition de termes pour enrichir une ontologie [81, 82, 29, 62].

Poibeau [29] a distingué deux types d'approches pour l'acquisition de classes de termes: les approches d'acquisition exogène et les approches d'acquisition endogène.

Les approches d'acquisition exogène se basent sur des ressources générales (ex. WordNet [110], Dictionnaire Intégral [29]) existantes pour certaines langues. La majorité des approches d'acquisition n'hésitent pas à utiliser ces ressources, parce qu'elles ont l'avantage d'être disponibles, immédiatement utilisables, indépendantes du domaine. Elles permettent d'acquérir des classes homogènes [29]. Cependant l'utilisation de ces ressources (et surtout WordNet) a été souvent critiquée pour les raisons suivantes :

- Le nombre élevé de sens pour un mot entraîne une ambiguïté dans l'acquisition [28].
- Elles sont abstraites, générales et elles encodent rarement les traits propres à un domaine visé.
- Les classes acquises peuvent ne pas comporter certains mots très fréquents dans le corpus.

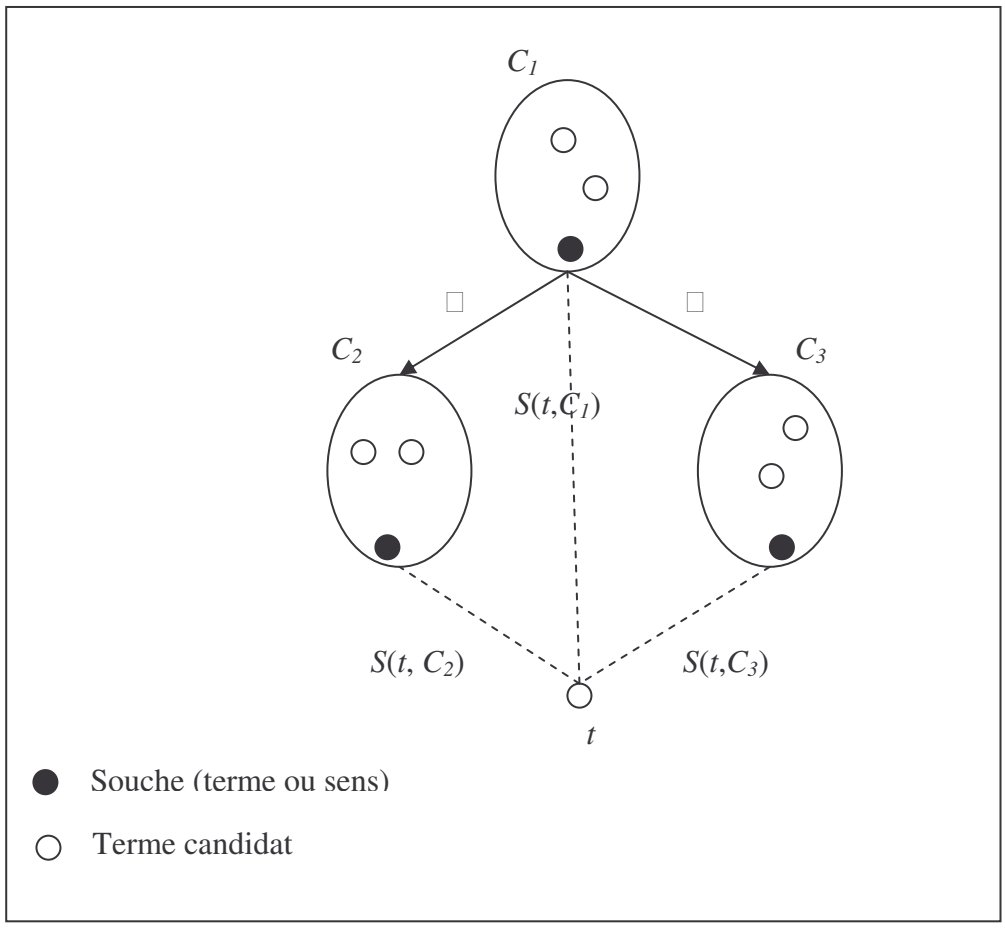


Figure 2.6: Association d'un terme *t* à une classe

Par conséquent, ces ressources sont souvent traitées comme des ressources incertaines [82]. L'association d'un terme *t* à une classe C_j la plus appropriée (cf. figure 2.6), se fait par une estimation de la similarité $S(t,C_i)$ entre *t* et chaque classe C_i définie pour l'application. Ces fonctions se basent essentiellement sur WordNet et surtout sur sa relation d'hyperonymie (*est-un*). Pour permettre le calcul de similarité, des termes t_{ij} appelés souche (*seed words* en anglais) ou sens des termes (c_{ij}) doivent être associés manuellement à chaque classe C_i [93]. La similarité $S(t,C_i)$ ($S(c,C_i)$) entre un terme *t* (ou un sens *c* de *t*) et une classe C_i est calculée entre le terme *t* (ou le sens *c*) et la souche des t_{ij} (ou c_{ij}) associée manuellement à la classe C_i. Lorsque la similarité est calculée entre termes, la classification est lexicale. Par conséquent, deux classes disjointes ne peuvent pas contenir un même terme et la similarité $S(t,t_i)$ peut être une fonction de la proportion des sens du terme *t* associés par la relation est-un relativement au sens du terme souche t_i de chaque classe C_i. La fonction $S_C(t, c_i)$ de Cimiano [83] permet une telle évaluation de la similarité :

$$S_C(t,t_i) = \max\left(\frac{\left|paths(senses(t), senses(t_i))\right|}{\left|senses(t)\right|}, 1\right)$$

où $paths(senses(t), senses(t_i))$ représente les chemins d'hyperonymie entre les sens du terme *t* et les sens du terme t_i.

La similarité sémantique $S(c_{jt},c_i)$ est calculée entre les sens c_{jt} du terme *t* et les sens souche c_i. Dans ce cas, deux classes disjointes ne peuvent pas contenir un même sens. De nombreuses fonctions de similarité sémantique ont été définies pour cette tâche de classification de termes ou pour la désambiguïsation du sens d'un mot. Budanitsky et Hirst [90] en présentent et en évaluent quelques unes. Un exemple de ce type de fonctions est la fonction $S_L(c_{jt},c_i)$ utilisée par Leacock et Martin [107] pour la désambiguïsation du sens des mots dans les textes en anglais par référence à WordNet.

$$S_L(c_{jt}, c_i) = -\log\left(\frac{len(c_{jt}, c_i)}{2 \times \max_{c \in WN}(depth(c))}\right)$$

où $len(c_{jt},c_i)$ est le plus petit chemin de c_{jt} à c_i dans WordNet (*WN*) et $depth(c)$= $len(root,c)$.

Les approches d'acquisition endogène constituent une sorte d'EI. Des patrons d'extraction sont utilisés pour repérer des relations binaires entre termes [81]. Quelquefois, le Web est utilisé comme une source certaine de connaissances pour l'extraction de ces relations [81, 95]. D'autres approches se basent sur la distribution des structures dans les différentes phrases. « L*e repérage des structures syntaxiques régulières permet de mettre en évidence des familles de mots apparaissant dans des contextes communs. Ces familles de mots servent à former des classes sémantiques* [97, 29, 96] ». Prenons, par exemple, la structure « *verbe préposition nom* ». Si cette structure est régulière dans un corpus (on la trouve dans le corpus avec le même *verbe* et la même *préposition*), alors les différents *noms* qui suivent la *préposition* sont rattachés à une même classe.

Les approches endogènes sont utiles pour acquérir les classes en l'absence de ressources générales. Selon Piobeau [29], ce type d'approches est bien adapté à la tâche d'analyse syntaxique. Mais, les classes ainsi obtenues ne sont pas complètement suffisantes. Leur pertinence est nuancée, elle dépend de l'affinité de regroupement, de la taille de corpus et de son homogénéité. Par conséquent, ces classes ne peuvent pas être utilisées directement dans une application, elles nécessitent un important travail de validation.

ASIUM [39, 29] est un système d'acquisition endogène. Il est semi-automatique (interactif). Il se base sur un corpus syntaxiquement annoté pour aider l'utilisateur à créer un ensemble de classes sémantiques et à enrichir une ontologie. En ajoutant des mots validés (souche) dans la classe à modéliser, les expérimentations menées avec ce type de système ont montré la supériorité de l'apprentissage supervisé sur l'apprentissage non supervisé [29]. Cardie [52] présente une approche d'acquisition supervisée des sens des mots à partir d'un corpus, un algorithme basé sur un arbre de décision est utilisé pour apprendre le contexte des mots. Hasttings [53] présente une autre approche (Camille) d'acquisition incrémentale et automatique à partir d'exemples.

Acquisition des règles d'extraction

Dans la section 2.1.2 nous avons présenté les différents types d'approches proposées et utilisées pour faciliter l'acquisition de ces règles et l'adaptation des systèmes d'extraction à d'autres applications.

Normalement, la généralisation des règles d'extraction exige des exemples positifs et des exemples négatifs pour diriger l'induction des règles. Dans un corpus d'entraînement, les annotations représentent les exemples positifs et le reste forme des exemples négatifs. Certaines approches [29, 34, 91, 93] se sont basées sur des classes de termes pour automatiser l'annotation sémantique des termes pertinents dans un corpus non annoté. D'autres approches, comme celles de Riloff [30] Riloff et Jones [93], Yangarber *et al.* [92], sont allées plus loin. Elles se sont basées sur une souche de patrons définis en arrière plan pour l'annotation automatique des passages candidats. Dans le cas de l'apprentissage supervisé [61, 58], à l'amorçage, la généralisation des règles nécessite un nombre réduit d'exemples manuellement annotés. Ensuite au fur et à mesure de la généralisation, le système teste les règles sur le corpus d'entraînement non annoté et les exemples les moins sûrs (les moins couverts et les plus informatifs) sont proposés à l'utilisateur pour l'annotation manuelle jusqu'à que les exemples incertains soient tous annotés.

Dans l'acquisition on peut distinguer aussi deux types d'approches : numérique et symbolique.

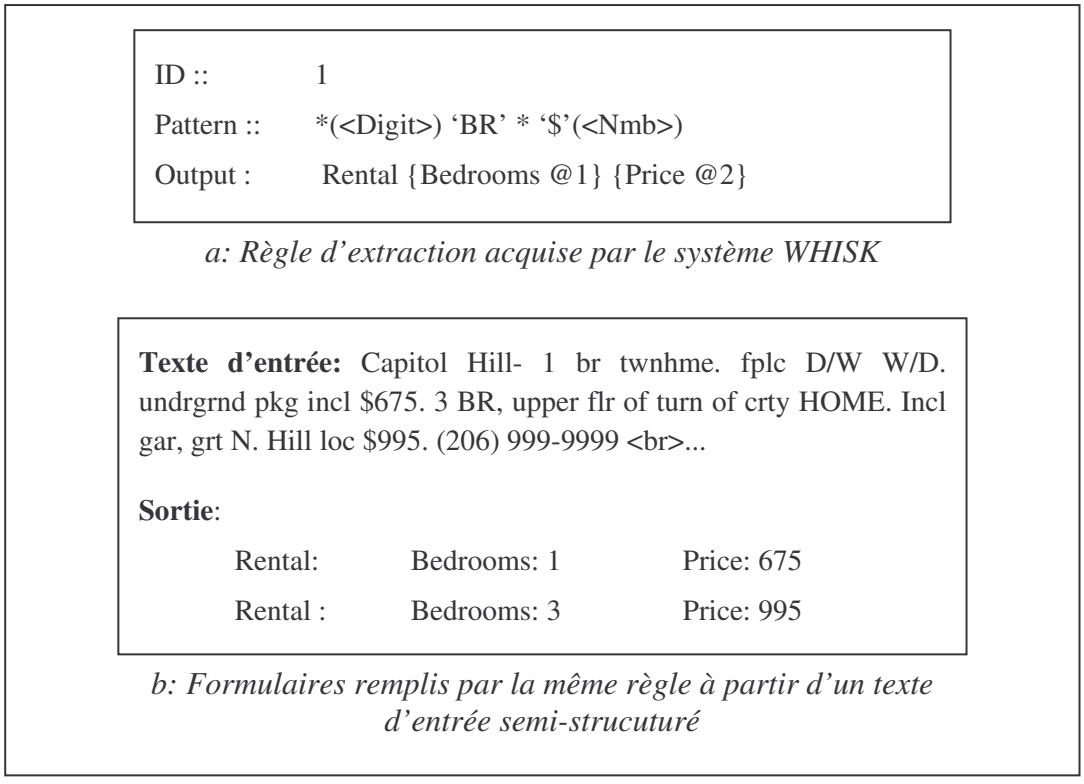
Les approches numériques [33, 87] construisent un modèle d'extraction statistique comme le modèle de Markov caché [100], en se basant sur la probabilité pour émettre une transition d'un état à l'autre dans le modèle.

Les approches symboliques construisent des modèles lisibles (et modifiables) par l'homme. Leurs règles sont définies par des concepts et des relations entre concepts. Ces approches se basent sur des arbres de décision [103], sur l'induction logique [98, 99, 102] ou sur une autre méthode d'apprentissage des relations (apprentissage relationnel) [101]. La généralisation de règles dans ces approches est souvent de type spécifique-à-général (*bottom-up*) et plus rarement de type général-à-spécifique (*top-down*).

CHILLIN [89], RAPIER [85, 94] sont des approches de génération combinant des méthodes d'induction logique *bottom-up* [98, 99] et *top-down* [102]. Les algorithmes SRV [63] et $(LP)^2$ [32] combinent des méthodes relationnelles et statistiques inspirées de (*Naïve Bayes*). La première est *top-down* et l'autre est *bottom-up*. CRISTAL [31] utilise une approche d'apprentissage relationnelle qui ressemble aux algorithmes d'induction logique de type *bottom-up*. Dans cette approche, Soderland *et al.* se basent sur une mesure de similarité entre règles pour les unifier dans une seule généralisation.

Les systèmes d'acquisition diffèrent entre eux par le nombre de champs (*slots*) du formulaire d'extraction qui peuvent être acquis en même temps. Les systèmes sont catégorisés en gros en *multi-solt* et *single-slot* [61].

Par exemple, le système WHISK [61] est un système d'acquisition *bottom-up* des règles *multi-slots*. Les règles ressemblent à des expressions régulières. A titre d'exemple, une règle vide de trois champs a la forme suivante "*(*)*(*)*(*)*". Les champs (*slots*) sont représentés par les étoiles entre parenthèses. Le reste des étoiles sont les délimiteurs de la phrase. Cette règle est vide, parce qu'une étoile signifie sauter jusqu'au prochain élément qui est aussi une étoile. Par contre, la règle de la figure 2.7.*a* lit une phrase contenant un chiffre suivi par la chaîne 'BR' et le caractère '$' suivi par un nombre. La ligne « Output » de la même figure indique que les champs lus doivent alimenter un formulaire appelé « Rental » (le premier nombre lu alimente le champ (*slot*) « Bedrooms » et le deuxième alimente le champ « Price » (cf. figure 2.7.*b*).



a: Règle d'extraction acquise par le système WHISK

b: Formulaires remplis par la même règle à partir d'un texte d'entrée semi-strucuturé

Figure 2.7: Exemple de règle d'extraction du système WHISK

Un exemple d'approche d'acquisition *single-slot* est le système AutoSlog [50]. Chaque règle d'extraction est acquise dans un cadre appelé « Concept Node » (cf. figure 2.8.*a*). Ce cadre définit des éléments pour la généralisation des règles comme le nom de la règle, un

déclencheur « trigger », le champ à extraire « Variable slots », des contraintes sur le contenu du champ « Constraints », et des conditions pour l'application de la règle « Enabling Conditions ». Par exemple, la règle présentée dans la figure 2.8.*a* se déclenche, dans les phrases contenant le mot « bombed ». Elle extrait le sujet des phrases passives et le remplit dans le formulaire, s'il appartient à la classe « physical-object ». « Constant Slots » dans cette règle définit une constante « bombing » qui impose d'extraire uniquement des événements de type « bombing ». AutoSlog se base sur treize patrons d'extraction pour acquérir ces cadres. La règle de la figure 2.8.*a* doit être généralisée par le patron d'extraction (<*subject*> *passive-verb*) pour traiter l'exemple présenté dans la figure 2.8.*b*.

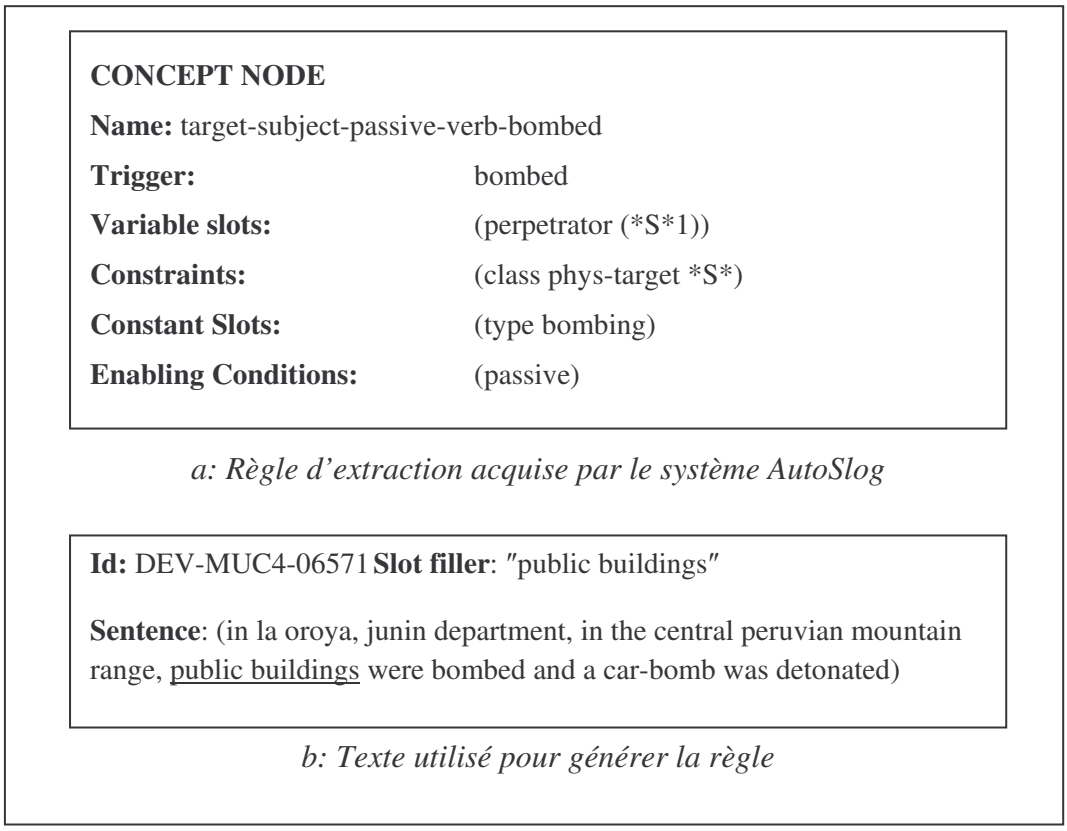


Figure 2.8: Exemple de règles d'extraction du système AutoSlog

Dans la majorité des systèmes d'acquisition, des cadres d'acquisition *multi-slot* ou single-*slot* sont définis [108]. Des cadres semblables à celui d'AutoSlog sont utilisés dans plusieurs systèmes comme « Concept Node » dans CRYSTAL [31], « FP-structure » dans PALKA [51], et « Egraph » dans HASTEN [36].

2.1.4 Tâches d'extraction d'information

Pour l'évaluation des systèmes d'extraction MUC-6 et MUC-7 (cf. § 2.1.5), l'EI a été séparée en cinq tâches : le repérage des EN (cf. § 2.1.4.1), la résolution des co-références (cf. § 2.1.4.2), le remplissage d'un formulaire d'éléments (cf. § 2.1.4.3), le remplissage d'un formulaire de relations (cf. § 2.1.4.4), le remplissage d'un formulaire de scénario (cf. § 2.1.4.5).

L'objectif de cette séparation est de permettre d'évaluer équitablement les systèmes d'extraction en se basant sur le type des informations qui peuvent être extraites : des noms, des relations de types attribut-valeur, des relations binaires entre entités ou entre événements. Cette séparation a permis, par conséquent, d'orienter la recherche vers les tâches les moins maîtrisées ainsi que de faire apparaître des approches d'extraction et des systèmes spécifiques pour chaque tâche, comme IdentiFinder^{TM} [33, 56] pour le repérages des EN et l'approche de Yangarber *et al.* [91] pour l'extraction des scénarios.

Les tâches d'extraction dépendent les unes des autres. Les tâches difficiles comme l'extraction des relations sont simplifiées, en se basant sur des connaissances acquises par les tâches maîtrisées. Par exemple, comme le repérage des entités nommées est la tâche la plus maîtrisée, la majorité des systèmes d'extraction commence par elle pour permettre ensuite de repérer les passages candidats.

2.1.4.1 Repérage des entités nommées (EN)

L'objectif de cette tâche est de repérer et d'annoter les noms des entités de type personne, entreprise et lieu. Cette tâche est reconnue comme étant *générique* dans la mesure où tous les textes font usage des noms propres et que leur repérage semble à priori reproductible. Le repérage et l'annotation de ces noms et des autres comme les expressions de temps et les unités numériques sont les plus faciles à faire dans l'ensemble des tâches d'extraction. Les scores réalisés dans les campagnes d'évaluation MUC (cf. § 2.1.4), montrent que cette tâche est aussi la plus maîtrisée parmi les tâches de l'extraction. Pourtant, pour certains domaines, elle est beaucoup plus complexe et ne donne pas un très bon résultat (ex. les corpus techniques [29]).

La réalisation de cette tâche ne demande pas un traitement linguistique complexe. Une segmentation superficielle des textes en groupes nominaux et verbaux est souvent utilisée

[54]. Les entités à extraire peuvent être ensuite repérées par des patrons à assortir localement aux groupes nominaux [54, 69, 33].

Tout d'abord, les EN (cf. figure 2.9) sont repérées par des indicateurs (amorces). Par exemple, les noms des personnes, peuvent être repérés :

- par leur titre : *M. Chirac* ;
- par un suffixe : *Snippty smith Jr.* ;
- par une initiale : *Humble T. Hopp* ;
- par une profession : *Le premier ministre Edouard Balladur*.

With heavy rains flooding northern <LOCATION> France </LOCATION>, a cartoon in L'Express this week shows <TRIGGER> Mr. </TRIGGER> <PERSON_NAME> Chirac </PERSON_NAME>, the <UNKNOWN> Gaullist </UNKNOWN>leader, trapped on a roof waiting to be rescued as <TITLE> Prime Minister </TITLE> <FIRST_NAME> Edouard </FIRST_NAME> <UNKNOWN> Balladur </UNKNOWN> nonchalantly walks away across the water.

Figure 2.9: Repérage des entités nommées (EN) dans un texte en langue naturelle [54]

Les noms d'organisations, comme Oracle dans la phrase « *Oracle Corp. in December had forecast third-quarter new software sales…* », sont souvent reconnus par des amorces terminales comme *Inc*, *Corp.*, *Corporation*, *Associates*, *Bank*…

Faute d'amorce, le repérage se base sur des dictionnaires spécifiques ou des ressources sémantiques (ontologies, réseaux sémantiques). Cependant, les dictionnaires disponibles ne sont pas toujours suffisants pour résoudre les conflits. Par exemple, dans la phrase « *Oracle in December had forecast third-quarter new software sales* », il n'y a pas une amorce devant ou derrière le mot Oracle permettant d'inférer le type de son entité. En revenant à une ressource générale (ex. WordNet 2.0), le mot oracle a plusieurs sens, mais aucun des sens présents ne permet d'inférer que *Oracle* est une organisation. Dans cet exemple, le conflit ne peut être résolu qu'avec une compréhension du texte, en mettant en relation plusieurs mots du texte : « *Oracle* » avec « *software* » et avec l'événement « *sale* ».

Les entités numériques et les dates sont souvent repérables par une analyse définie par des critères formels [29, 84].

Ensuite, la segmentation en groupes nominaux et verbaux permet de vérifier et délimiter l'étendue de chaque annotation (cf. figure 2.10) en utilisant une grammaire de segmentation spécifique (cf. figure 2.5).

Enfin, des patrons linguistiques sont appliqués localement sur les passages des entités repérées [29, 54] pour extraire les entités pertinentes. Cette tâche est utilisée dans certaines approches [29], pour l'enrichissement automatique des dictionnaires. Les entités inconnues au départ (encapsulés entre les balises <UNKNOWN> et </UNKNOWN> dans la figure 2.9) et reconnues par la suite sont enrichies dans un dictionnaire spécifique.

With heavy rains flooding northern <LOCATION> France </LOCATION>, a cartoon in L'Express this week shows <PERSON> Mr. Chirac </PERSON>, the Gaullist leader, trapped on a roof waiting to be rescued as <TITLE> Prime Minister </TITLE> <PERSON> Edouard Balladur </PERSON> nonchalantly walks away across the water.

Figure 2.10: Amélioration de l'annotation des entités repérées

Plusieurs types d'approches et de systèmes sont développés pour cette tâche [79], fondés sur des règles définies manuellement ou par apprentissage (par des règles d'induction logiques [76], par un arbre de décision [77] par un modèle numérique [33, 78]) supervisé ou non supervisé [80].

2.1.4.2 Résolution de co-références

L'objectif de cette tâche est d'identifier les variantes d'une même entité pour permettre de regrouper les informations à son sujet. Les variantes peuvent être une abréviation, un acronyme, une variation orthographique, un synonyme, un hyperonyme ou un autre type comme une anaphore pronominale. Ces variantes sont recherchées dans les groupes nominaux d'un texte [28].

La résolution de co-références inter-phrases fait nécessairement appel à des ressources sémantiques pour mettre en relation les termes du texte. WordNet est là aussi une des

premières ressources utilisées [71]. Dans le cas de l'anaphore pronominale, la résolution est un problème du TALN difficile à résoudre [28, 70], parce qu'elle demande une compréhension du discours. Ainsi, beaucoup de connaissances linguistiques sont nécessaires pour définir les règles de résolution et pour améliorer la performance actuelle de cette tâche [72]. Jusqu'à présent, l'acquisition de ces connaissances se fait par apprentissage car cette approche a réussi à mettre en place des techniques de résolution plus simples que celles du TALN avec une performance raisonnable [71].

2.1.4.3 Extraction des attributs (*template element*)

Le but de cette tâche est de repérer des informations descriptives des EN et de les alimenter dans un formulaire spécifique, c'est le formulaire d'éléments. Les informations descriptives cherchées sont des valeurs pour des attributs définis dans le formulaire et dépendant des entités. A titre d'exemple, pour la campagne d'évaluation MUC-7, un formulaire était conçu pour des entités de types organisation, personne et artéfact [73]. Les attributs et leurs valeurs autorisées pour l'extraction des caractéristiques d'une entreprise étaient *NAME*, *ALIAS*, *ORG_DESCRIPTOR*, *TYPE* avec les valeurs (*GOVERNMENT*, *COMPANY*, *OTHER*), *RG_LOCALE et ORG_COUNTRY*. Pour une personne, ces attributs étaient (*NAME*, *ALIAS et TITLE*) [74].

Dans les systèmes d'extraction, il n'y a pas de modules dédiés à cette tâche. Cette tâche est traitée comme une tâche simplifiée de l'extraction de relations (cf. § 2.1.4.4) associant un attribut à une valeur. L'extraction des relations attribut-valeur commence par le repérage des EN et une résolution de co-références pour identifier les passages candidats contenant des informations descriptives des entités [74, 54]. Des classes de termes associant sémantiquement les entités à leurs attributs et une analyse morphologique et lexicale pour résoudre les ambiguïtés sont des éléments très utiles dans le repérage des attributs. Enfin les relations attribut-valeur d'une entité sont repérables par des patrons linguistiques mettant en relation syntaxique des classes sémantiques de termes. Les scores donnés pendant la conférence MUC-7 (cf. figure 2.12) montrent que cette tâche est plus maîtrisée que celle de l'extraction de relations (87% contre 76%).

2.1.4.4 Extraction de relations (*template relation*)

L'objectif de cette tâche est d'extraire des relations binaires (faits) entre les EN et de les remplir dans un formulaire spécifique appelé le formulaire de relation. A titre d'exemple, trois relations binaires étaient demandées pour la campagne d'évaluation MUC-7 : « produit-de » entre un artéfact et une organisation, « employé_de » entre une organisation et une personne, « situation_de » entre un lieu et une organisation [73].

L'extraction des relations se focalise sur les relations intra-phrase [28]. Les relations intra-phrases sont très compliquées et nécessitent une compréhension complète du discours. Les relations intra-phrase sont de deux types :

1. Des relations repérables par des marqueurs de surface comme (A est B ; A est appelé B) ;

2. Des relations dépendant des rôles syntaxiques des entités dans la phrase (sujet, verbe, objet direct, objet indirect,…). L'extraction de ce type de relation par une approche fondée sur le TALN est très difficile, puisque les règles d'extraction doivent prendre en compte de nombreuses variations :

    - Des variations syntaxiques des éléments de la relation visée: active « *X* produit *Y* », passive « *Y* est produit par *X* » et nominative « la production de *Y* par *X* » ;

    - Des variations stylistiques provenant de l'utilisation de termes sémantiquement similaires « *Y* est mis au point par *X* » sont possibles. Comme pour l'extraction des événements, les variations stylistiques sont surmontées par un regroupement des termes dans des classes sémantiques.

Dans le domaine biologique Ray et Craven [87] utilisent un modèle de Markov caché pour apprendre à extraire ce type de relation binaire comme la relation (protéine, localisation) représentant la localisation d'une protéine dans la cellule.

2.1.4.5 Extraction des événements (*template scenario*)

Cette tâche est l'ultime objectif d'un système d'extraction. Elle doit permettre de remplir un formulaire prédéfini de scénario par des informations descriptives sur des événements relatifs à ce scénario (cf. figure 2.1). Ces informations une fois remplies dans le formulaire

forment une sorte de compte-rendu sur le scénario permettant de répondre à des questions ‘qui, quoi, quand, comment...’ pour chaque événement. Cette tâche est reconnue comme la plus difficile des tâches de l’extraction. Son meilleur score réalisé pendant l’évaluation de MUC-6 (cf. tableau 2.12) était moins de 57%. Dans cette tâche, l’information à extraire peut être distribuée sur plusieurs phrases. L’extraction nécessite de la repérer et de la combiner. Pour cette combinaison, des informations implicites doivent être explicitées. Un traitement superficiel du texte est peu utile pour cette tâche qui nécessite une compréhension complète du texte en langue naturelle. Les nuances grammaticales et chronologiques perturbent beaucoup l’extraction et doivent être prises en compte dans les règles d’extraction [54]. Prenons, par exemple les deux phrases suivantes :

- *Sam était président, Harry lui a succédé* et
- *Sam sera président, il suit Harry*.

Ces phrases ne fournissent pas la même information sur la succession des personnes sur le poste du président. L’information qu’on extrait de la première phrase est que Sam n’est pas président tandis que la deuxième phrase nous informe exactement de l’inverse.

Actuellement, l’extraction des relations inter-phrases est considérée comme très difficile et est souvent abandonnée. L’extraction des événements se fait alors par un traitement local et superficiel pour permettre l’application des patrons linguistiques impliquant des classes sémantiques de termes définissant les événements et les verbes concernés. La figure 2.11 présente un extrait des patrons d’extraction définis manuellement pour le système FASTUS [49, 68]. Dans cette figure les classes de termes à extraire sont encapsulées entre ‘<’ et ‘>’.

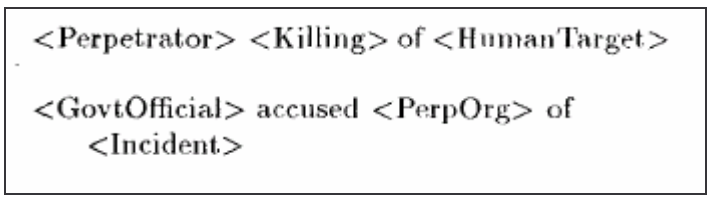


Figure 2.11: Extrait des patrons d’extraction utilisés dans le système FASTUS

Pour l’acquisition automatique, Yangaber [91] utilise un modèle de Markov caché et se base sur des classes sémantiques de termes et sur l’analyse syntaxique des exemples (phrases) pertinents pour acquérir son modèle.

2.1.5 Evaluation des systèmes d'extraction

2.1.5.1 Evaluation objective

Un cadre d'évaluation et de comparaison des systèmes d'extraction d'information a été proposé par les campagnes d'évaluation américaines MUC (*Message*[2] *understanding conferences*). Des métriques objectives de type boîte noire [29], ont été proposées et affinées pendant les conférences MUC-2 (1989) et MUC-3 (1991). Ces métriques permettent de mesurer le bruit (information erronée extraite) et le silence (information pertinente non repérée) générés par un système. Les métriques de base utilisées sont la précision *P* et le rappel *R* [54, 66, 29]:

La précision *P* en IE permet de mesurer la quantité des réponses correctes $N_{correctes}$ parmi les réponses totales retenues par le système $N_{réponses}$:

$$P = \frac{N_{correctes}}{N_{réponses}}$$

Le rappel *R* mesure la quantité des réponses correctes $N_{correctes}$ parmi les réponses clés $N_{clés}$ ($N_{clés}$ sont les réponses références remplies par l'organisateur).

$$R = \frac{N_{correctes}}{N_{clés}}$$

Quelques fois une autre mesure supplémentaire appelée *F-mesure* est utilisée pour combiner et faire la synthèse entre la précision *P* et le rappel *R*. Cette mesure est établie comme suit :

$$F - mesure = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$$

Le coefficient β permet de privilégier soit les systèmes qui ont une meilleure précision *P* ou ceux qui ont un meilleur rappel *R*. Souvent, β est fixé à 1 pour donner un poids équivalent à la précision et au rappel. La formule obtenue est appelée *P&R* (*la combinaison de P et R*) :

---

[2] Message signifie un texte court et informatif [29].

$$P \& R = \frac{2 \times P \times R}{P + R}$$

2.1.5.2 Evaluation subjective

L'évaluation subjective mesure l'ergonomie générale du système, la présence ou non d'experts, de corpus annotés, de ressources extérieures, la rapidité et le pourcentage des tâches correctement réalisées par un groupe d'utilisateurs représentatifs.

Ce type d'évaluation est rarement fait parce qu'il est difficile à mettre en place. Des utilisateurs d'un certain niveau d'expérience n'ont pas toujours la disponibilité pour effectuer le test nécessaire [29].

2.1.5.3 Récapitulatif des conférences MUC

Dans les conférences MUC, les participants disposaient d'une description détaillée de l'information à extraire avec un ensemble de documents (corpus d'entraînement) et les formulaires à remplir à partir de ces documents. Après un délai de 1 ou 6 mois donné pour l'adaptation de leurs systèmes, les participants obtiennent de nouveaux textes (corpus de test) et retournent à l'organisateur de la conférence les formulaires remplis automatiquement.

Pendant la conférence MUC-4 (1992), les systèmes sont évalués sur un corpus de récits d'attentats en Amérique latine. A partir de ce corpus les participants devaient extraire des informations concernant la date, le lieu, le type, la victime et l'auteur du crime et les remplir dans une instance d'un schéma prédéfini. Dans l'évaluation (cf. figure 2.12), les principaux systèmes ont du mal à dépasser le 60 % de la mesure *P&R*.

Pour la conférence MUC-5 (1993) deux domaines (création de groupes d'entreprises et microélectronique) et deux langues naturelles (l'anglais et le japonais) sont proposés. Le formulaire établi permettait l'utilisation des champs de type pointeur [66]. Cette conférence était la plus difficile. Plusieurs participants ont échoué dans l'adaptation manuelle de leurs systèmes aux objectifs fixés. Le meilleur score obtenu (cf. figure 2.12) était moins que 65% pour *P&R*.

L'évaluation des systèmes dans la conférence MUC-6 (1995) était basée sur quatre tâches d'extraction : l'annotation des EN, la résolution de co-référence, le remplissage d'un

formulaire d'éléments, le remplissage d'un formulaire de scénarios [74]. Les meilleurs scores enregistrés pour ces tâches sont présentés dans la figure 2.12.

Evaluation\Tasks	Named Entity	Coreference	Template Element	Template Relation	Scenario Template	Multilingual
MUC-3					R < 50% P < 70%	
MUC-4					F < 56%	
MUC-5					EJV F < 53% EME F < 50%	JJV F < 64% JME F < 57%
MUC-6	F < 97%	R < 63% P < 72%	F < 80%		F < 57%	
MUC-7	F < 94%	F < 62%	F < 87%	F < 76%	F < 51%	
Multilingual						
MET-1	C F < 85% J F < 93% S F < 94%					
MET-2	C F < 91% J F < 87%					

Legend: R = Recall P = Precision F = F-Measure with Recall and Precision Weighted Equally
E = English C = Chinese J = Japanese S = Spanish
JV = Joint Venture ME = Microelectronics

Figure 2.12: Meilleurs scores obtenus pendant les campagnes d'évaluation MUC

La conférence MUC-7 (1998) a marqué la fin de ces campagnes. Dans cette dernière, une autre tâche a été ajoutée à celles évaluées en MUC-6, c'est le formulaire de relations. La performance des systèmes était autour des scores enregistrés dans la conférence précédente MUC-6. Pour le formulaire de relations le meilleur score était inférieur à 76% [67].

2.2 Systèmes de question-réponse SQR

2.2.1 Approches générales des systèmes de question-réponse

Comme présenté au début de ce chapitre, question-réponse QR est la technique consistant à extraire à partir d'une grande collection de documents comme le Web des réponses à des questions posées par l'utilisateur. Cette technique possède des points de similitude avec l'EI dans la mesure où il s'agit de réduire le champ d'extraction à des types spécifiques d'informations. Cependant un traitement linguistique beaucoup plus complexe est nécessaire pour «comprendre» l'objet de la question [28].

B. Grau [40] a catégorisé les SQR en deux types d'approches :

- Le premier type est orienté compréhension de textes. Son idée est de construire une représentation sémantique profonde des textes en langues naturelles. Le système Qualm [41] est un exemple de ce type d'approches.

- L'autre type est plus récent et orienté EI. Il s'agit de concevoir la tâche de QR comme le corollaire d'une application d'EI dans laquelle le système doit répondre à des demandes d'informations précises, en représentant l'objet de la question comme un formulaire d'extraction. Pour cette représentation, des relations sémantiques entre les entités de la question et son objet sont extraites, en se basant sur une analyse syntaxique locale (segmentation en groupes nominaux et verbaux) de la question. Ensuite, le repérage du document candidat pour extraire la réponse fait appel à une technique d'appariement « requête-document » connue en recherche documentaire (RD). Enfin, l'extraction d'une réponse devient une technique d'EI, en remplissant le formulaire de la question à partir de ce document candidat (cf. PowerAnswer [47]).

2.2.2 Architecture générale d'un système de question-réponse

Selon Grou [40, 59], un SQR (cf. figure 2.13) comporte trois modules principaux. Ce sont le module d'analyse de questions, le module de traitement de documents et le module d'extraction de la réponse.

2.2.2.1 Module d'analyse de questions

La tâche principale du module d'analyse de questions est d'identifier la catégorie de la question [41, 42] pour déterminer le type de la réponse attendue et son focus. Le focus de la question est l'objet sur lequel la réponse doit focaliser, par exemple dans la question « Quelle est la formule chimique du dioxyde de soufre ? », le focus est le dioxyde de soufre.

Ce module peut avoir aussi pour tâche de rechercher des autres reformulations possibles de la question afin d'augmenter la capacité des autres modules à identifier la réponse [28].

Pour réaliser ses tâches, ce module effectue des traitements linguistiques syntaxiques et sémantiques de la question. Le traitement syntaxique est pour identifier la catégorie de la question et son focus. Ensuite, le traitement sémantique peut être utilisé pour identifier des

spécifications du focus, pour reconnaître le type de la réponse attendue, et pour identifier les relations sémantiques entre les entités de la question nécessaires pour localiser la réponse. Ce traitement peut être fait, en se basant sur un réseau sémantique général comme WordNet ou une ontologie spécifique au domaine d'application [109]. A titre d'exemple, pour la question «quel volcan a détruit la ville de Pompei ? », une ressource sémantique générale comme WordNet peut être utilisée pour déduire « volcano » comme un type général de la réponse attendue (cf. QALC [59]).

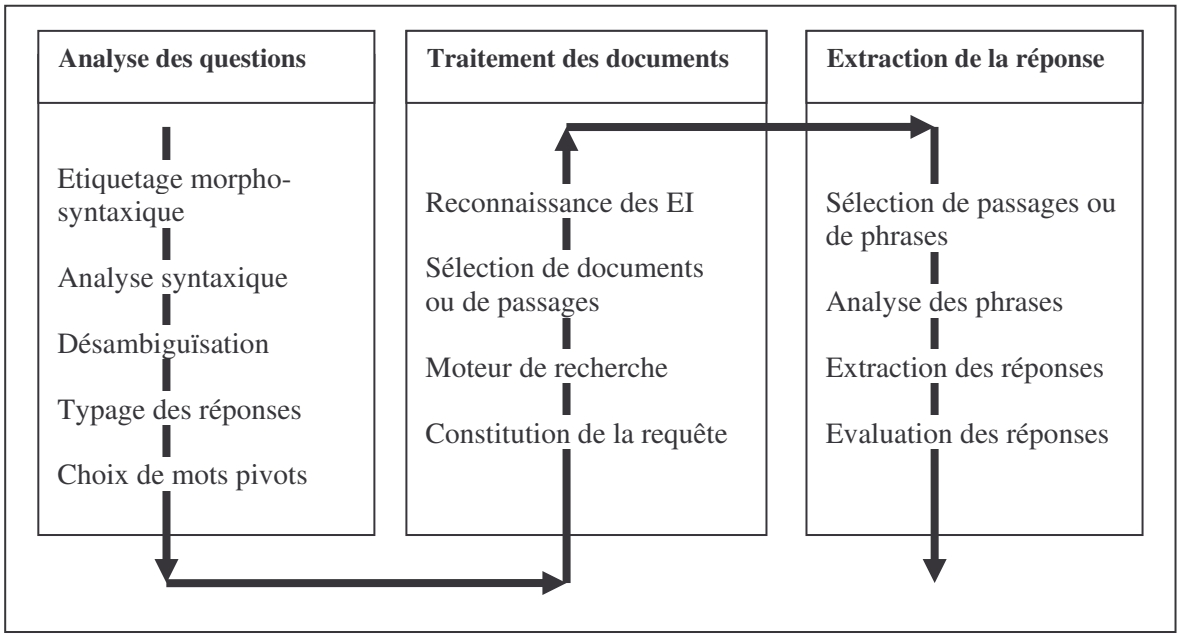


Figure 2.13: Architecture générale des systèmes de question-réponse

2.2.2.2 Module de traitement des documents

L'objectif de ce module est de localiser les passages candidats contenant la réponse. Ces passages peuvent être localisés par une technique de recherche documentaire RD, en s'appuyant sur la densité des mots dans le passage [28]. Ensuite, dans certaines approches comme QALC [11], ces passages peuvent subir un traitement plus poussé afin de rechercher toutes les variations de termes constitutifs [35, 38] de la requête par lemmatisation[3] et par expansions[4] par le moyen des outils comme Fastr [43] et TermWatch [37].

[3] La lemmatisation d'un mot est le processus d'associer un mot à sa forme canonique (forme de base).
[4] L'expansion est définie par l'ajout de nouveaux éléments dans les deux positions grammaticales possibles dans un syntagme nominal SN : centre (tête) et modifieur.

Dans les approches linguistiques [28], des patrons représentant la question et permettant de modifier certains constituants (prépositions, adjectives, adverbes,…) peuvent être automatiquement définis et utilisés pour localiser les passages candidats. Ces patrons peuvent aller jusqu'à la définition d'une forme affirmative à la place d'une forme interrogative afin de simplifier le repérage des passages contenant une réponse. Par exemple, la question « Quand *est né Lyndon B. Johnson* » peut être reformulée comme « <*focus*> <*verbe principe*>… ». Ce qui donne le patron « *Lyndon Johnson est né le …* ».

2.2.2.3 Module d'extraction de la réponse

L'objectif de ce module est de choisir parmi les passages candidats le passage contenant la réponse ou d'extraire une réponse exacte. Le cas le plus facile est celui où le type de la question est une entité nommée. Dans ce cas, il suffit de proposer la chaîne contenant l'entité. Dans les autre cas, l'extraction de la réponse est plus difficile, parce qu'elle peut être formulée de plusieurs façons stylistiques et grammaticales. Comme dans l'EI, l'extraction de la réponse se fait par des patrons linguistiques [28]. Soubottin [44] en présente une approche linguistique manuelle. L'énumération et la définition de tous les partons nécessaires pour l'extraction de la réponse est plus difficile que dans l'EI. Des patrons sont à prévoir pour chaque type de questions. Certains types comme les questions de relation nécessitent des patrons très complexes pour mettre en évidence des relations sémantiques entre entités. Lin et Panel [45] ont proposé une approche endogène non supervisée inspirée de celle de l'acquisition endogène des classes de termes (cf. § 2.1.3.2). Dans cette approche l'acquisition des règles d'inférence de relations est une extension de l'hypothèse distributionnelle des mots [97]. Dans cette extension, on suppose que si deux relations syntaxiques sont régulières dans un corpus syntaxiquement analysé (elles apparaissent souvent avec les mêmes mots), alors elles sont sémantiquement similaires.

Une autre approche est issue des travaux de Moldovan *et al.* [46]. Cette approche est basée sur la compréhension des textes ; une représentation syntaxique de chaque passage candidat est appariée avec la représentation syntaxique de la question par un raisonnement abductif. Ce raisonnement est un processus de génération des hypothèses en présence des faits. Il permet d'attribuer un score à chaque hypothèse (réponse candidate) et ensuite de choisir l'hypothèse (la réponse) la plus sûre. Dans leurs nouveaux travaux [47], ils développent un système utilisant le démonstrateur Cogex pour transformer la question et le

passage candidat sous forme logique (logique de premier ordre). Le système se base aussi sur une série d'axiomes pour modéliser des connaissances du monde réel. Ces axiomes sont dérivés de plusieurs sources : WordNet, axiomes ontologiques, axiomes linguistiques manuellement définis et axiomes temporels issus de la base de connaissances Sumo [45]. Dans ce système, le score attribué à chaque réponse candidate est sa similarité syntaxico-sémantique avec la question calculée par les modules du système.

2.2.3 Evaluation des systèmes de question-réponse

La campagne internationale d'évaluation des systèmes QR la plus connue est TREC « *Question Answering Track* ». Lors de son édition TREC 11 en 2002, les évaluateurs ont constaté la difficulté de la production d'une réponse exacte aux questions par rapport à la sélection d'un passage contenant une réponse [48]. Dans l'édition TREC 12, en 2003, deux tâches ont été soumises aux systèmes. La première tâche s'appelle la tâche de passage « *passage task* » et demande un passage de 250 caractères comme réponse à des questions factuelles (*factoid questions*). La deuxième s'appelle la tâche principale « *main task* » et demande de fournir une réponse exacte pour trois types de questions : question factuelle, question demandant en réponse une liste d'entités, question demandant une définition. L'évaluation des questions de type définition est la plus difficile des trois, parce qu'il ne s'agit pas de juger la justesse de la réponse mais d'attribuer des points pour les parties de la réponse correspondant à des concepts présents dans la question. La difficulté est donc de définir des limites pour l'étendue d'une réponse pour permettre à des juges humains de donner des jugements approchés [28]. Par ailleurs, partant des réponses données par les systèmes aux questions de définition, on a l'impression que les questions sont interprétées comme des questions factuelles [28, 48]. Par exemple, aucun des passages retournés (cf. figure 2.14) comme réponse à la question « *what is a golden parachute* ? » [5] ne contenait une définition du terme « *golden parachute* ».

L'édition TREC 14 (2005) a soumis deux tâches aux systèmes : la tâche principale et la tâche de relation « *relationship task*». Dans la tâche principale, une nouveauté est introduite, c'est la question factuelle imbriquée. Dans ce type de questions (cf. figure 2.15) une réponse

[5] Golden parachute est une somme colossale payée à un grand patron d'entreprise en cas de départ pour le mettre à l'abri des besoins. Voir une autre définition aussi sur le site : http://fr.wikipedia.org/wiki/Parachute_en_or

extraite pour une question est nécessaire pour extraire une réponse aux questions suivantes. Ainsi, le système qui a de la difficulté pour trouver la réponse exacte pour une de ces questions, aura plus de chance pour trouver une bonne réponse pour les suivantes.

Extrait 1. *If he left, on leaving, O'Neill would be able to collect a golden parachute package providing three years of salary and bonuses, stock and other benefits.*

Extrait 2. *The big payment that Eyler received in January was intended as a « golden parachute ».*

Extrait 3. T*he arrangement which includes iucrative stock options, a hefty salary, and a « golden parachute » if Gifford is fired…*

Figure 2.14: Exemple de réponses données par un système à la question de définition « *what is a golden parachute* »

Dans la tâche de relation [57], relation est définie comme la capacité d'une entité d'influencer une autre, par le sens ou par la motivation d'agir (ex. *What impact did the Rat Pack have on the rise of the Las Vegas tourism industry?*).

Q_1. *Who was the first Imam of the Shiite sect of Islam?*

Q_2. *Where is his tomb?*

Q_3. *When did he die?*

Figure 2.15: Exemple de questions imbriquées de *TRAC 14*

Les scores obtenus par les 15 meilleurs systèmes restent modestes. Les systèmes les plus performants à l'édition TREC 14 sont les deux systèmes de Harabaigu *et al.* [47] : PowerAnswer pour la tâche principale « *main task* » et Palantir pour les questions de relation. Les scores obtenus pour ces systèmes sont 71% pour les questions factuelles, 46% pour les questions de liste (dont la réponse est une liste d'entité nommées) et 22% pour les autres types de questions (relationnelles et de définition).

En plus de TREC, d'autres campagnes d'évaluation des SQR sont intervenues comme CLEF au niveau européen et EQUER au niveau national français. Leurs méthodes d'évaluation et les scores obtenus sont présentés dans [28].

2.3 Bilan

Cette étude de l'état de l'art nous conduit à constater une similarité technique des deux approches étudiées. Toutes deux se basent principalement sur le repérage dans un texte en langue naturelle des entités nommées. Ces entités sont définies par le formulaire d'extraction pour un système d'extraction ou identifiées dans la question pour un système de question-réponse. Ces deux approches mettent ces entités en relation syntaxique et sémantique pour extraire l'information pertinente. Par rapport aux systèmes d'extraction où les entités sont définies par le formulaire, les systèmes de question-réponse présentent une difficulté supplémentaire qui est l'analyse de la question pour définir les entités à localiser et les entités à extraire.

Chapitre 3

Principes de l'approche proposée

et problèmes posés

Dans ce chapitre, nous présentons l'approche que nous proposons. Nous explicitons les problèmes de mise en œuvre de cette approche et nous montrons que les solutions actuellement disponibles ne conviennent pas pour différentes raisons. L'approche consiste à concevoir un système d'aide à l'innovation (SAI) basé sur une ontologie d'innovation O_{ino}. Dans un domaine de connaissances Δ, ce système peut être utilisé pour générer une base de connaissances et l'enrichir par des exemples de résolution inventive de problèmes. Un système d'extraction (SE) est intégré dans le système SAI pour extraire ces exemples à partir des textes en langue naturelle LN. Un système de résolution de problèmes (SRP) emploie cette base pour aider l'utilisateur (expert, étudiant, chercheur, …) à la résolution inventive de ses problèmes.

3.1 Approche proposée

Comme déjà présentée dans le chapitre 1, l'innovation est une activité globale. Les problèmes d'innovation et leur résolution concernent la majorité des domaines de connaissances. Néanmoins, les efforts déployés pour généraliser cette activité sont encore peu nombreux.

- Description de la créativité et des problèmes d'innovation comme un acte général et indépendant des domaines.
- Elaboration de méthodes et d'algorithmes généraux pour l'aide à la résolution des problèmes d'innovation.
- Elaboration de bases de connaissances et d'exemples d'innovation.

- Développement d'une infrastructure qui exploite le contenu de ces bases de connaissances pour la résolution inventive des problèmes d'innovation.

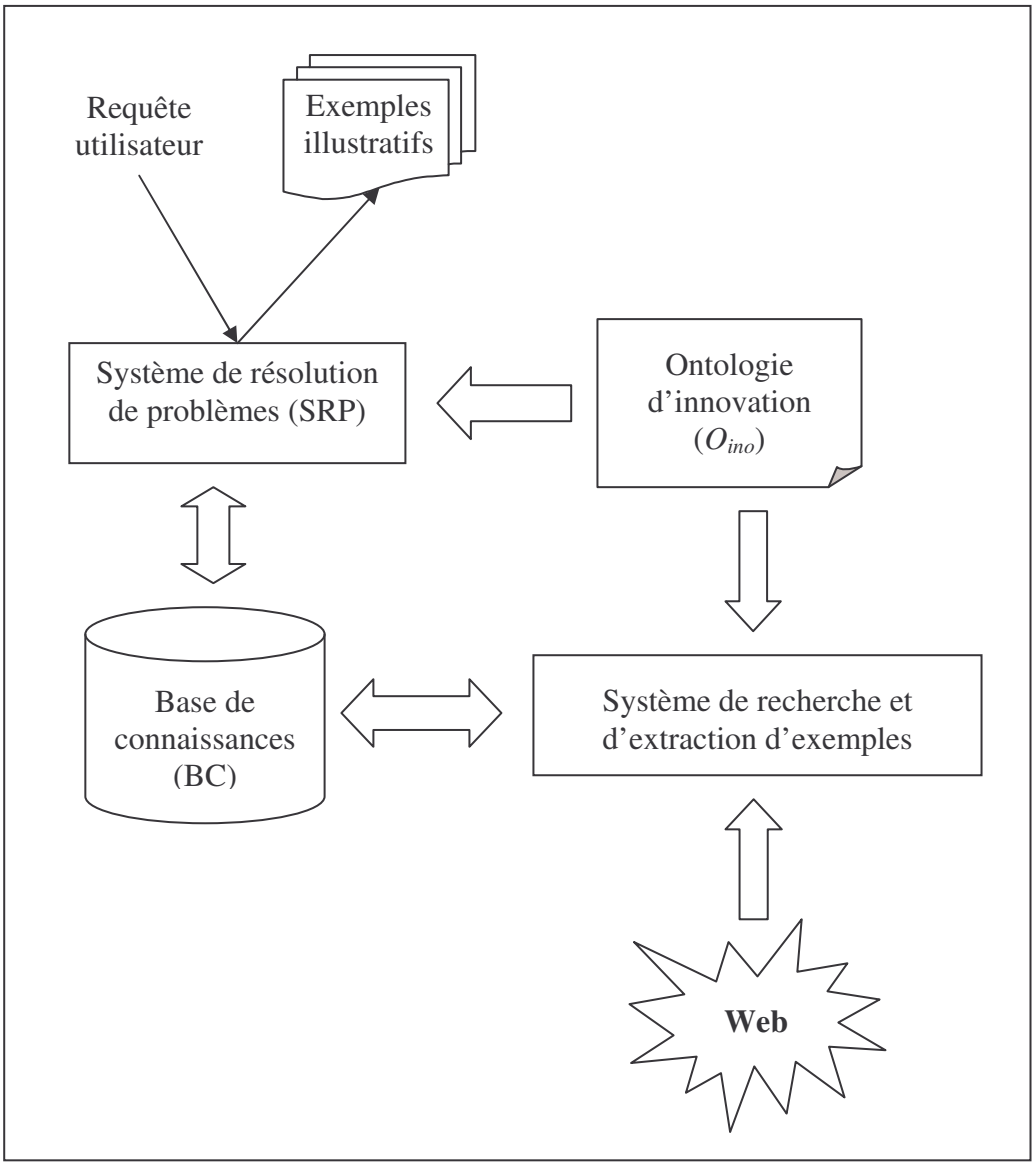


Figure 3.1: Synoptique du système d'aide à l'innovation SAI proposé

Dans ce travail, nous proposons de résoudre ces différents problèmes en déployant une ontologie d'innovation O_{ino} (cf. chapitre 4). Un expert dans un domaine de connaissances se base sur cette ontologie pour définir une base de connaissances BC et des opérateurs d'innovation pour son domaine. Un système de résolution de problèmes (SRP cf. § 3.2) se greffe sur cette ontologie, sur la base BC et sur des exemples d'application des opérateurs d'innovation afin d'aider l'utilisateur à une résolution inventive de ses problèmes.

Puisque les bases de connaissances d'innovation actuelles sont des collection de textes en langues naturelles sur le Web, un système de recherche et d'extraction d'exemples (cf. § 3.3) est aussi proposé pour extraire les exemples et enrichir la base de connaissances BC. Le but de cet enrichissement est de faciliter le raisonnement et le repérage des solutions par le système SRP ou par l'utilisateur dans la base BC. Les deux systèmes (SRP et le système de recherche et d'extraction) sont intégrés dans un système général appelé le système d'aide à l'innovation (SAI) (cf. figure 3.1).

Dans la suite de ce chapitre, nous présentons plus en détail le fonctionnement de ces systèmes ainsi que les approches possibles pour les implémenter.

3.2 Système de résolution de problèmes

Selon Altshuller (cf. § 1.1), l'inférence d'une solution inventive consiste tout d'abord à identifier le problème, ensuite à le généraliser, puis à caractériser des solutions génériques et enfin d'interpréter ces solutions pour en tirer des solutions spécifiques au problème. La méthode d'innovation TRIZ formule des solutions génériques comme des opérateurs d'innovation. En s'appuyant sur des exemples d'application de ces opérateurs, la résolution d'un problème posé consiste donc à identifier les opérateurs d'innovation correspondant au problème, puis à extraire des exemples de ces opérateurs à partir d'une base de connaissances, et enfin inférer une solution pour le problème.

Ainsi, pour effectuer une résolution inventive et automatique des problèmes, il nous faut les éléments suivants:

1. Une identification des opérateurs d'innovation dans les bases de connaissance.
2. Une représentation formelle de ces opérateurs par une ontologie d'innovation O_{ino}.
3. Des exemples d'opérateurs identifiables dans les bases de connaissances d'innovation.
4. Un mécanisme du raisonnement basé sur les éléments précédents pour la résolution automatique des problèmes.

Les bases de connaissances d'innovation sont des collections des textes en langues naturelles sur le Web. Pour faciliter l'accès aux exemples dans ces textes par l'utilisateur et par le système SRP, des métadonnées[6] décrivant les exemples doivent en être extraites.

Puisque l'extraction manuelle de ces métadonnées est très difficile, nous proposons un système d'extraction (cf. § 3.3) automatique pour les mémoriser dans la base de connaissances BC. Dans cette base, les exemples sont sémantiquement représentés en s'appuyant sur ces métadonnées. On peut facilement les utiliser pour le développement du système SRP. On peut les utiliser aussi pour annoter automatiquement des termes pertinents qui représentent des concepts (type d'opérateur, type de ressource) définies dans l'ontologie.

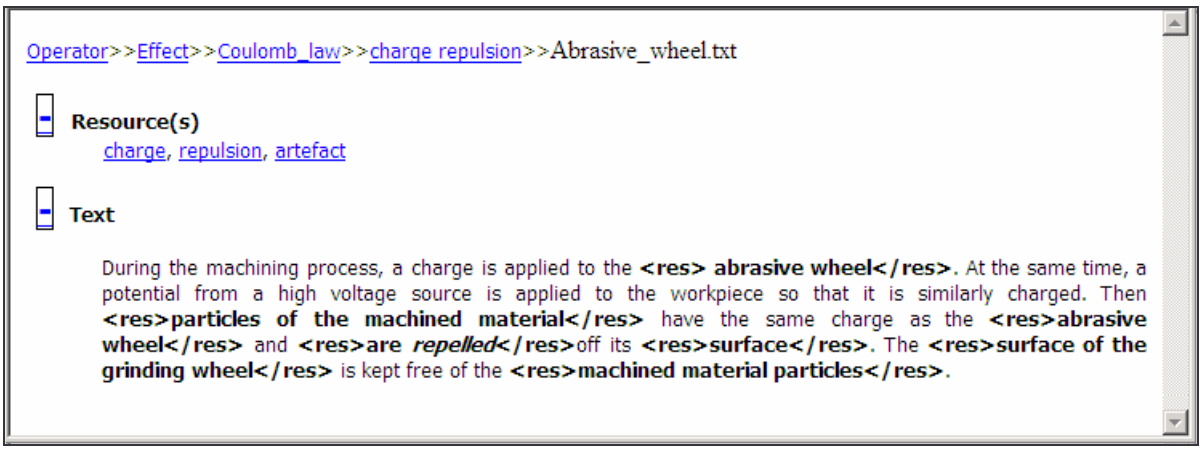


Figure 3.2: Interface utilisateur

Ce travail s'est principalement focalisé sur la conception et le développement de ce système d'extraction. A cause des nombreuses difficultés rencontrées, une résolution entièrement automatique n'a pas pu être retenue. Cependant à travers une interface appropriée (cf. figure 3.2), un utilisateur peut faire évaluer les exemples et les métadonnées extraits. Quand un exemple dans la base de connaissances est affiché sur cette interface, des annotations dépendant de la requête de l'utilisateur sont dynamiquement insérées dans le texte d'origine. La base BC est interrogeable par des hyperliens intégrés dans l'interface. Ces hyperliens représentent les types des opérateurs et les ressources définis dans la base de connaissance. En cliquant sur ces liens, l'utilisateur peut facilement extraire de cette base les exemples associés.

---

[6] Une métadonnée est une donnée sur une donnée. Dans les systèmes d'information, c'est une information "compréhensible" par la machine (cf. http://www.w3.org/Metadata/).

3.3 Système de recherche et d'extraction automatique des exemples

Ce système (cf. figure 3.3) est composé de deux modules. Le premier module est un système de recherche documentaire SRD. Sa tâche est de repérer sur le Web des textes pertinents du domaine de connaissances Δ et de ses opérateurs représentés dans la base BC. Le deuxième module est un système d'extraction des exemples SE. Ce système récupère les textes retournés par le système de recherche SRD et en extrait des exemples d'opérateurs d'innovation. Après avoir traité tous les textes retournés, il peut redemander au système SRD de lancer une nouvelle recherche sur le Web.

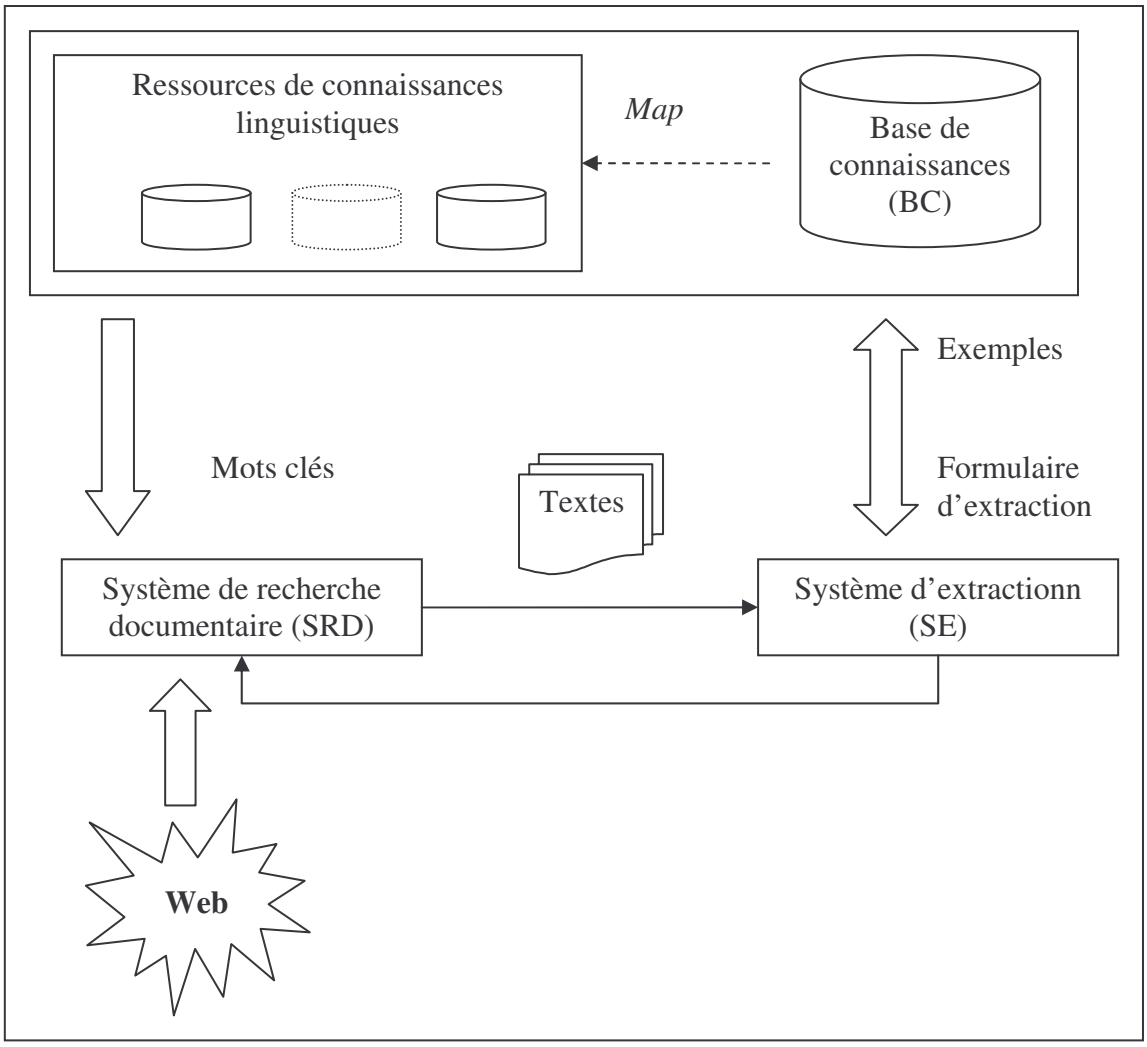


Figure 3.3: Système de recherche et d'extraction proposé

Un mapping *Map* (cf. chapitre 4) entre les entités définies dans la base de connaissances BC et les classes sémantiques des termes et des mots définis dans des ressources linguistiques est nécessaire pour extraire les exemples des textes.

3.3.1 Système de recherche documentaire

Dans une grande collection documentaire comme le Web, les documents et les ressources de données sont très hétérogénes et variés relativement à leur type (des textes, des images, des bases de données, d'autres types de ressources) et à leur domaine de connaissances. Pour extraire les textes pertinents à un domaine Δ, ce module compose automatiquement des requêtes adressées à un moteur de recherche sur le Web (Google, Yahoo, …). Pour cette composition des requêtes, des mots clés sont nécessaires. Ce module doit être capable de les identifier dans la base de connaissances BC ou dans des ressources linguistiques à travers le mapping *Map*.

Le problème qui se pose est celui de l'amorçage du système de recherche et d'extraction lorsque la base BC est vide d'exemples.

Une première solution consiste à effectuer l'amorçage par le système d'extraction SE. Dans ce cas, l'utilisateur doit sélectionner les documents nécessaires pour l'amorçage. L'inconvénient de ce type d'amorçage est que beaucoup de textes pertinents sur le web peuvent ne pas être repérés s'ils emploient des termes différents de ceux alimentés dans la base de connaissances à partir de ces documents.

Une autre solution consiste à effectuer l'amorçage à partir du système de recherche. Dans ce cas, le système de recherche doit employer le mapping *Map* pour sélectionner les mots clés. Dans cette approche, les mots clés peuvent être très nombreux. Ils peuvent être utilisés afin de composer une infinité de requêtes de recherche. Les requêtes peuvent être mieux spécifiées, en utilisant des mots clés positifs (qui doivent exister dans les documents) et négatifs (qui ne doivent pas exister dans les documents) à partir des ressources linguistiques. Le seul problème est de trouver des ressources linguistiques bien adaptées au domaine d'application Δ.

Dans les deux approches, un mécanisme de raisonnement pour la sélection des mots clés est nécessaire. Elle doit permettre de composer des requêtes valides par lesquelles un moteur de recherche retourne le maximum des documents pertinents et le minimum du bruit

(documents non pertinents). Pour cet objectif, nous avons adopté une approche de recherche proposée par Santamarìa *et al.* [121], en nous basant sur le mapping *Map* instancié avec le réseau sémantique WordNet (cf. chapitre 4).

3.3.2 Système d'extraction des exemples

Comme déjà présenté dans le chapitre 2, ce système doit prendre en entrée un texte en langue naturelle (cf. figure 3.4.*a*) et alimenter en sortie le formulaire d'extraction (cf. figure 3.4.*b*). Le contenu de ce formulaire représente un exemple de résolution innovante à insérer dans la base de connaissances BC (cf. figure 3.3). Cette solution est définie par un opérateur d'innovation dans un domaine de connaissances. Un expert du domaine doit définir un tel formulaire pour chaque opérateur de son domaine.

La figure 3.4.*b* présente un exemple des motifs qu'on souhaite extraire et enrichir dans la base BC pour un opérateur du type effet nommé loi de Coulomb (*Coulomb_law*) (figure 3.4.*a*). En plus du lien vers le texte défini par le motif *Link*, d'autres motifs définissent les ressources de l'exemple et son opérateur (*Operator*). Les ressources définies dans ce formulaire sont de deux catégories. La première est la catégorie des ressources sur lesquelles l'opérateur est réalisé, autrement dit les objets de l'opérateur. Ces objets sont extraits par les motifs *Resource* et *Type* identifiés à partir de l'ontologie et du mapping. Ces deux motifs ne sont pas connus à l'avance et ils peuvent correspondre à plusieurs occurrences dans le formulaire. L'autre catégorie de ressource est constituée par les ressources recommandées par l'opérateur pour résoudre le problème. Le motif *Cause* définit les ressources causes de l'effet et le motif *Improve* définit la ressource améliorée par l'opérateur. Pour notre opérateur loi de Coulomb ces types sont respectivement *charge* et *repulsion* et sont spécifiés par l'expert. Le formulaire est alors extrait par le système à partir du texte (cf. figure 3.4.*a*). Pour chaque type, on peut remplir dans le formulaire de 1 à n termes (un ou plusieurs mots significatifs du domaine). Le terme "*charge*" est associé au motif *Cause* et un terme "*repel*" est associé au motif *Improve*. Deux termes objets "*surface of the granding wheel*" et "*material particle*" du type *physical_object* sont aussi extraits et remplis dans le formulaire.

Le formulaire présenté dans la figure 3.4.*b* représente un cadre pour des opérateurs de type effet. Il s'agit d'une relation cause-conséquence entre ressources où la conséquence est la ressource améliorée. D'autres cadres peuvent être définis pour les deux autres types d'opérateurs. Les motifs *Improve*, *Operator*, *Link* sont commun à tous les types. La définition

des autres cadres consiste à spécifier d'autres motifs dans le formulaire et à leur associer un type de l'ontologie d'innovation. Le motif *Type* dans le formulaire est nécessaire pour alimenter la base de connaissances et pour permettre son exploitation sémantique dans l'aide à la résolution des problèmes.

Abrasive wheel load prevention

During the machining process, a charge is applied to the abrasive wheel. At the same time, a potential from a high voltage source is applied to the work piece so that it is similarly charged. Then particles of the machined material have the same charge as the abrasive wheel and are repelled off its surface. The surface of the grinding wheel is kept free of the machined material particles.

a: Exemple de l'effet loi de Coulomb « Une force d'attraction existe entre deux charges de signes différentes. Une force de répulsion existe entre deux charges similaires »

Link: "Abrasive_wheel.txt"	**Type:** Example
Improve: "repel"	**Type:** repulsion
Resource: "surface of grinding wheel"	**Type:** physical_object
Resource: "material particle"	**Type:** physical_object
Cause: "charge"	**Type**: charge
Operator: Coulomb_low	**Type** effect

b: Formulaire de solution rempli à partir de l'exemple ci-dessus

Figure 3.4: Exemple illustratif de la tâche d'extraction exigée pour un domaine d'innovation

En entrée, le système d'extraction doit se baser sur l'ontologie O_{ino}, sur les opérateurs définis dans la base de connaissances BC ou sur leurs formulaires et sur le mapping *Map*. Un texte peut définir un exemple pour un ou plusieurs opérateurs, le système doit donc pouvoir extraire plusieurs formulaires d'un même texte.

En revenant au formulaire de la figure 3.4, on constate que les informations à extraire représentent des relations sujet-action-objet (i.e. *material_particle-repel-grinding_wheel*). Par conséquent, nos formulaires ressemblent au formulaire de scénarios dans les systèmes d'extraction. Dans ces formulaires, on a toujours des événements à repérer et à extraire.

Cependant, les systèmes d'extraction d'information ne sont pas adaptables à notre application pour des raisons présentées dans la section suivante. Ainsi, la réalisation de cette tâche nécessite un système spécifique.

Pour repérer ces relations sujet-action-objet, nous avons essayé de traiter le problème d'extraction par une approche linguistique (cf. § 3.4.2). Cependant, les analyseurs syntaxiques existants (cf. annexe E) ne nous sont pas apparus pertinents pour repérer ces relations. Il faudrait les associer avec des règles d'extraction difficiles à spécifier. Pour ces raisons et d'autres présentées dans la section suivante, nous avons abandonné cette voie. L'apprentissage aussi n'était pas envisageable à cause du nombre élevé de formulaires potentiels à apprendre et à cause de l'absence d'un corpus d'entraînement (cf. § 3.4.1).

La voie qui nous est apparue la plus appropriée (cf. chapitre 4) repose sur une approche sémantique adaptée à notre compétence. Cette approche (cf. chapitre 4) se base sur une analyse lexicale, sur une mise en relation sémantique des mots apparus dans le même contexte et sur des ressources sémantiques (WordNet, l'ontologie O_{ino}). Cette approche est inspirée des approches de désambiguïsation des sens des mots dans les textes. L'avantage de cette approche est qu'elle nous a permis de développer deux tâches en parallèle. La première est le développement et l'enrichissement du modèle sémantique de l'ontologie ; la deuxième est le développement du module d'extraction et son implémentation.

Cette approche peut être illustrée sur le formulaire et le texte présenté dans la figure 3.4 :

- On définit d'abord les ressources (*artefact*, *repulsion*, *charge*) dans la base de connaissances BC comme instances d'une classe Ressource dans l'ontologie O_{ino}.

- On associe par le mapping *Map* les termes nécessaires à ces ressources. Soient *Map*(*artefact*)={"*particle*", "*wheel*", "*surface*"}, *Map*(*charge*)={"*charge*"} et *Map*(*repulsion*)={"*repel*"}.

- On fait une analyse lexicale du texte permettant d'identifier le lemme (forme canonique) et la catégorie lexicale de chaque mot.

- On associe chaque terme à sa ressource.

- Des termes peuvent être associés par le mapping *Map* à des ressources ayant ou non des relations sémantiques (hyperonyme/hyponyme, partie/tout,…). Ainsi, une approche de désambiguïsation dans l'application doit être définie pour remplir adéquatement le formulaire. Par conséquent, un traitement linguistique superficiel peut être mis en place pour filtrer les informations et les formulaires pertinents.

3.4 Adaptation d'un système d'extraction à notre application

A travers une étude comparative des systèmes d'extraction, il apparaît que les systèmes d'extraction existants ne sont pas adaptables à notre application. Ces systèmes comme présenté dans le chapitre 2 sont fondés sur deux approches : le TALN et l'apprentissage. Ces approches posent divers problèmes pour notre application. Les approches d'apprentissage (cf. § 3.4.1) nécessitent un corpus d'entrainement qui est très délicat à élaborer. Les approches TALN ne sont pas adaptables en raison de la nature des motifs et des arguments à extraire (cf. § 3.4.2) et aussi à cause des difficultés de mise au point des règles d'extraction.

3.4.1 Adaptation d'une approche fondée sur l'apprentissage

L'apprentissage peut être supervisé ou non-supervisé (cf. § 2.1.2).

L'adaptation d'une approche supervisée à notre application nécessite l'élaboration d'un corpus d'entraînement, ce qui consomme beaucoup de temps et d'efforts de la part de l'expert. Pour l'extraction des solutions inventives, des formulaires d'extraction doivent être définis pour chacun des opérateurs du domaine de connaissances. Afin d'apprendre ces formulaires pour les opérateurs d'innovation définie dans la méthode TRIZ par exemple, un expert doit préparer des dizaines de textes d'entraînement pour chaque opérateur. Dans TRIZ des milliers d'opérateurs sont définis (cf. annexe B). En nous basant sur un système d'apprentissage, il nous faudrait plusieurs milliers de textes pour l'entraînement. De plus, des textes suffisamment représentatifs sont difficiles à trouver et pour certains opérateurs peuvent ne pas exister.

Pour l'apprentissage non-supervisé, des ressources sémantiques (ontologie, réseau sémantique) associé à des termes sont indispensables. L'association de ces termes à l'ontologie d'innovation peut être envisagée manuellement ou automatiquement à partir de ressources linguistiques générales (i.e. WordNet). Toutefois, l'acquisition des classes à partir

de ce genre de ressources est risquée dans l'extraction. Nous avons remarqué que des relations sémantiques entre termes utiles pour notre application ne sont pas représentées dans WordNet. Dans notre application, ces relations sont nécessaires pour l'extraction des exemples de certains opérateurs d'innovation définis dans la méthode TRIZ. Par exemple, des opérateurs d'innovation dans TRIZ dépendent de l'état solide des ressources. Ces ressources doivent être en principe représentées par des ressources hyponymes de la ressource *physical_object*. Si on cherche les *synsets* (cf. chapitre 4) ayant des relations d'hyponymie avec les *synsets* du mot "*solid*" dans WordNet, on ne trouve qu'un petit nombre d'hyponymes. Bien que la ressource *electron* soit un hyponyme de la ressource *electric_charge*, WordNet ne permet pas d'inférer cette relation. Aussi l'action *cavitation* signifiant l'injection de bulles d'air dans un liquide n'existe pas non plus. Dans certains travaux (cf. § 2.1.3.2), on utilise des fonctions de similarité pour repérer des mots sémantiquement similaires. De toute façon il faut disposer de relations sémantiques entre les mots dans le réseau WordNet. Sinon, WordNet est peu utile.

Des classes de termes mappés aux ressources des opérateurs d'innovation sont indispensables. Dans le chapitre suivant nous présentons une solution pour l'acquisition semi-automatique de ces classes à partir de WordNet.

3.4.2 Adaptation d'une approche fondée sur le TALN

Comme présenté dans la section 3.3, nos formulaires d'extraction peuvent s'apparenter au formulaire de scénario dans l'extraction d'information. Un formulaire de scénario (cf. § 2.1.4.5) décrit un événement par son sujet, son action, son objet, son lieu, son temps, etc. Un formulaire d'exemple décrit une fonction réalisée par ses ressources sujettes ou sa fonction cause, son action et ses ressources objets.

Malgré la ressemblance entre ces deux types de formulaires, les motifs de base sont d'une nature différente ainsi que leurs techniques d'extraction. Dans l'extraction d'information, ces motifs sont des entités nommées EN (cf. § 2.1.4.1). Elles sont souvent repérables par des patrons et des indicateurs linguistiques locaux. Dans notre application, leur repérage peut être intéressant. Mais elles ne représentent pas les motifs de base auxquels nous nous intéressons. Comme présenté dans la section précédente, les motifs qui nous intéressent sont les ressources (cf. § 1.2.3) définies dans les formulaires et décrits dans l'ontologie O_{ino}. Selon la méthode TRIZ, elles sont les éléments existant dans l'environnement (entités physiques, énergies, leur propriété cf. § 1.2.3) et nécessaires pour réaliser une fonction dans un système.

Les indicateurs à utiliser pour repérer ces ressources peuvent être des termes ambigus ayant plusieurs sens potentiels. A titre d'exemple, dans le texte de la figure 3.4.*a*, les termes *charge*, *abrasive wheel* et *material particles* représentent des ressources pour l'opérateur loi de Coulomb. Cependant si un de ces termes (comme *charge*[7]) a plusieurs sens le repérage de ces termes n'est pas suffisant pour l'interpréter comme une ressource. Ainsi le repérage des ressources dépend du contexte. Des relations sémantiques avec des autres entités, des indicateurs linguistiques, ou bien encore des règles d'extraction doivent être utilisés pour résoudre cette ambiguïté dans les textes. Dans notre approche (cf. chapitre 4), nous nous sommes aidés des opérateurs pour identifier le contexte et résoudre les ambiguïtés.

Dans l'extraction d'information, après le repérage des entités nommées EN le reste des tâches (cf. § 2.1.3) extraient des attributs, des relations entre ces entités et avec des événements prédéfinis. Dans notre application, après le repérage des ressources, des solutions inventives doivent être extraites. Cette tâche exige des tâches supplémentaires comme la mise en relation de contradictions (par les principes inventifs cf. § 1.2.13.1), la mise en relation cause-conséquence (par les effets cf. § 1.2.13.3) des ressources. Ces tâches supplémentaires ne sont pas effectuées dans les systèmes d'extraction et nécessitent une solution spécifique.

Comme il n'est pas dans notre objectif de développer un système de TALN spécifique à notre approche, nous avons essayé de voir comment un système existant, Pinocchio [55], pouvait être adapté. Pinocchio propose un environnement pour faciliter la spécification des ressources de connaissances et des règles d'extraction.

Pinocchio fournit un ensemble d'outils d'extraction d'information comme un analyseur syntaxique, des éditeurs, des compilateurs, des débuggeurs et des traceurs afin de faciliter la modification et la mise au point des ressources de connaissances (i.e. lexique, base de connaissances, hiérarchie, règles d'extraction syntaxiques et sémantiques). Il comporte aussi des navigateurs des formulaires et des structures de données extraits.

Pour faciliter la modification de ces ressources de connaissances, tous les modules de Pinocchio prennent en entrée la même structure de données (*token char*). Cette structure de

---

[7] Charge a plusieurs sens linguistiques et ses sens ne représentent pas tous des ressources. Par exemple, dans la phrase «his charge was deliver a message », charge veut dire mission et ne correspond pas à une ressource (c. WordNet).

données se présente sous la forme d'un graphe orienté (cf. figure 3.5). Ses sommets sont des objets élémentaires T_i (*tokens*) représentant abstraitement des éléments lexicaux a_i d'un texte *t* et ses arêtes sont des relations binaires définies dans trois structures associées à chaque sommet T_i :

- $[T_i]_{dep}$: relations syntaxiques associant a_i à un autre *token* dans *t*.
- $[T_i]_{feat}$: une structure représentant les informations syntaxiques et sémantiques nécessaires pour combiner a_i à un autre *token*.
- $[T_i]_{lf}$: une forme logique afin de donner une interprétation pour la combinaison de a_i à un autre *token*.

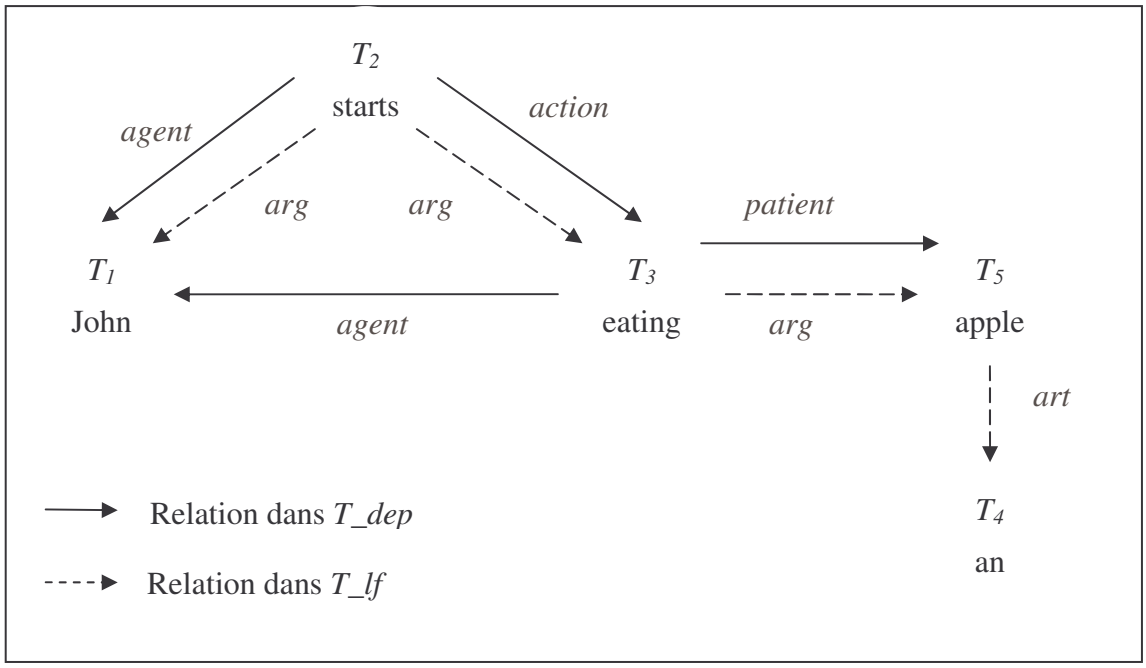


Figure 3.5: Graphe utilisé par Pinocchio pour l'extraction d'information. Ce graphe associe les éléments lexicaux de la phrase « *John starts eating an apple* » par des relations de $[T_i]_{dep}$ et $[T_i]_{lf}$

Les modules de Pinocchio fonctionnent sur le graphe par le moyen de règles d'extraction organisées en séquence de cascades. Ces règles accèdent, construisent et modifient le graphe. La forme générale d'une règle (cf. figure 3.6 et [112]) est un triplet: $< \gamma\alpha\delta, \Gamma_T, \Gamma_A >$ où

- $\gamma\alpha\delta$ est le patron de la règle dont α représente le centre et ne peut pas être vide, γ et δ représentent le contexte et peuvent être vides.

- Γ_T est la partie condition de la règle. C'est un ensemble de prédicats booléens définis sur le patron.

- Γ_A est l'action de la règle, c'est un ensemble d'opérations élémentaires sur le centre.

PATRON	**CONDITION**	**ACTION**	**ENTREE**
T_1	$[T_1]_{feat}.cat$=NP	*Dependant*($[T1]_{dep}$,$[T_3]_{dep}$) $[T1]_{feat}.subcat.int\text{-}arg = [T_3]_{feat}$ $[T_1]_{lf}.patient = [T_3]_{lf}.head$	"*the issue*"
T_2*	$[T_2]_{feat}$.cat=Adjunct		
T_3	$[T_3]_{feat}$.cat=PP $[T_3]_{feat}=[T_1]_{feat}$.subcat.int-arg		"*of bonds*"

Figure 3.6: Règle pour reconnaître [*the issue of bonds*]$_{np}$

Ce type de règles qui intègre des contraintes syntaxiques et sémantiques dans l'extraction est bien adapté à notre besoin pour résoudre l'ambiguïté du sens des termes. Cependant nous ne pouvons pas utiliser Pinocchio, pour plusieurs raisons :

- Tout d'abord, la mise au point des ressources de connaissances est manuelle dans Pinocchio. Cette tâche reste une tâche longue et difficile à faire. Nous voulons aussi la réduire pour permettre à un expert d'adapter le système à son domaine. Chaque modification dans le domaine d'application peut nécessiter une modification sur l'ontologie ou sur la base de connaissances et par conséquent sur les règles d'extraction de Pinocchio.

- Ensuite, dans notre situation, on n'est pas capable de définir le contexte syntaxique à partir duquel on doit extraire des éléments sémantiquement définis dans l'ontologie d'innovation. Les opérateurs, les ressources et leurs relations sont des entités sémantiques. Elles peuvent être interprétées de multiples façons par des personnes différentes pour des domaines différents. Le terme *charge* dans l'opérateur loi de Coulomb peut représenter aussi une particule (électron ou proton), ou plus généralement un objet sur lequel l'action charge est réalisée ou encore une action. Les textes en langue naturelle reflètent ces interprétations. Mais le mot *charge* n'est pas suffisant pour définir la bonne interprétation, il faut aussi connaître le contexte. En anglais, ce mot *charge* peut être un verbe ou un nom. Par conséquent il peut jouer

plusieurs rôles syntaxiques dans une phrase. Ces rôles ne nous aident pas pour déterminer l'interprétation. Il faut utiliser aussi les mots qui existent dans son contexte. Dans les phrases du texte de la figure 3.4, le terme *charge* est respectivement un sujet d'une phrase passive, une action (participe passé), un objet. Mais sémantiquement, en nous basant sur la loi du Coulomb, ces trois rôles syntaxiques présentent sémantiquement la même ressource qui produit la fonction de répulsion entre la roue (*wheel*) et les particules (*particle*). C'est donc bien le sujet de la fonction réalisée.

Ac magnetic field changes voltage accross coil

An ac magnetic field is applied parallel to the axis of a coil. As the field alternates, the *changing* magnetic flux induces an electric field in the turns of the coil. The induced field forces electrons to move, producing the induced current. The work the induced field does to move a unit positive charge around a closed contour is called the induced electromotive force (emf).

Figure 3.7: Opérateur d'innovation de type effet

- Enfin, les règles d'extraction ne permettent pas de résoudre l'ambiguïté au niveau des opérateurs. Par le mapping *Map*, des classes de termes doivent être associées au concept de l'ontologie (ressource et action cf. chapitre 4). Par conséquent, en nous basant sur les opérateurs de la méthode TRIZ, un terme peut apparaître dans plusieurs classes de termes. Il est alors mappé à plusieurs concepts et il participe à plusieurs rôles de différents opérateurs. Dans l'opérateur présenté dans la figure 3.7, le champ magnétique est une ressource utilisée pour produire un champ électrique. Ainsi le champ magnétique représente le sujet et le champ électrique est l'objet. Pour d'autres opérateurs ces rôles sont inversés. À cause de cette ambiguïté, un même texte peut être extrait comme exemple pour un opérateur mais aussi pour d'autres opérateurs. La clé de l'extraction des exemples des opérateurs d'innovation est la résolution de cette ambiguïté à tous les niveaux (terme, ressource, opérateur). On ne voit pas comment on peut utiliser Pinocchio pour cette désambiguïsation qui nécessite des informations provenant de plusieurs phrases dans le texte. De plus les opérateurs correspondent aux arêtes dans le graphe de Pinocchio. La résolution de l'ambiguïté entre opérateurs nécessite la mise au point de règles d'extraction

sémantiques sur ces opérateurs candidats. Pinocchio ne permet pas de mettre en œuvre ce genre de règles, qui sont des contraintes sur les arêtes du graphe et non pas sur les sommets.

3.5 Extraction des exemples comme triplets (sujet, action, objet)

Puisque l'extraction des exemples de résolution des problèmes d'innovation consiste à réaliser des fonctions, et puisqu'une fonction est une relation entre un sujet, une action et un objet, nous avons pensé à extraire des triplets (ressource, action, objet) des textes et de les comparer avec des triplets similaires représentant les opérateurs dans la base de connaissances.

Ce principe de solution est inspiré d'une approche proposée par Tsourikov *et al.* [111] pour la recherche et la classification des documents par rapport à une requête utilisateur en langue naturelle. Cette approche est brevetée et enregistrée au nom de la compagnie productrice du logiciel Goldfire innovator (cf. § 1.3.1). Dans cette approche, les documents candidats sont tout d'abord sélectionnés, en comparant leurs mots clés avec ceux de la requête. Les mots clés sont repérés par la fréquence de leur apparition dans le texte. Ensuite, des triplets composés d'un sujet *S*, d'une action *A* et d'un objet *O* sont extraits des phrases des textes et comparés avec des triplets similaires extraits de la requête. Suite à cette comparaison, une valeur de pertinence de 1 à 10 est donnée à chaque texte candidat.

Dans cette approche, Tsourikov *et al.* extraient les triplets (*S*,*A*,*O*) par des règles d'extraction spécifiques. L'extraction nécessite plusieurs étapes de traitement (cf. figure 3.8) pour accumuler les informations nécessaires à l'application des règles. Après l'étape d'extraction des triplets, une étape de normalisation intervient pour faciliter la comparaison de ces triplets entre la requête et les textes candidats.

Pour utiliser cette approche, il nous faut une représentation sémantique des opérateurs d'innovation comme des triplets (*S*,*A*,*O*) à la place des requêtes utilisateurs.

Le système développé par Tsourikov *et al.* étant breveté, il était nécessaire de trouver une autre approche pour l'extraction des triplets (*S*,*A*,*O*). Nous avons pensé utiliser un analyseur

syntaxique pour extraire des triplets (sujet, verbe, objet) à la place des triplets (*S*,*A*,*O*) [8]. Au final nous avons abandonné cette idée pour plusieurs raisons:

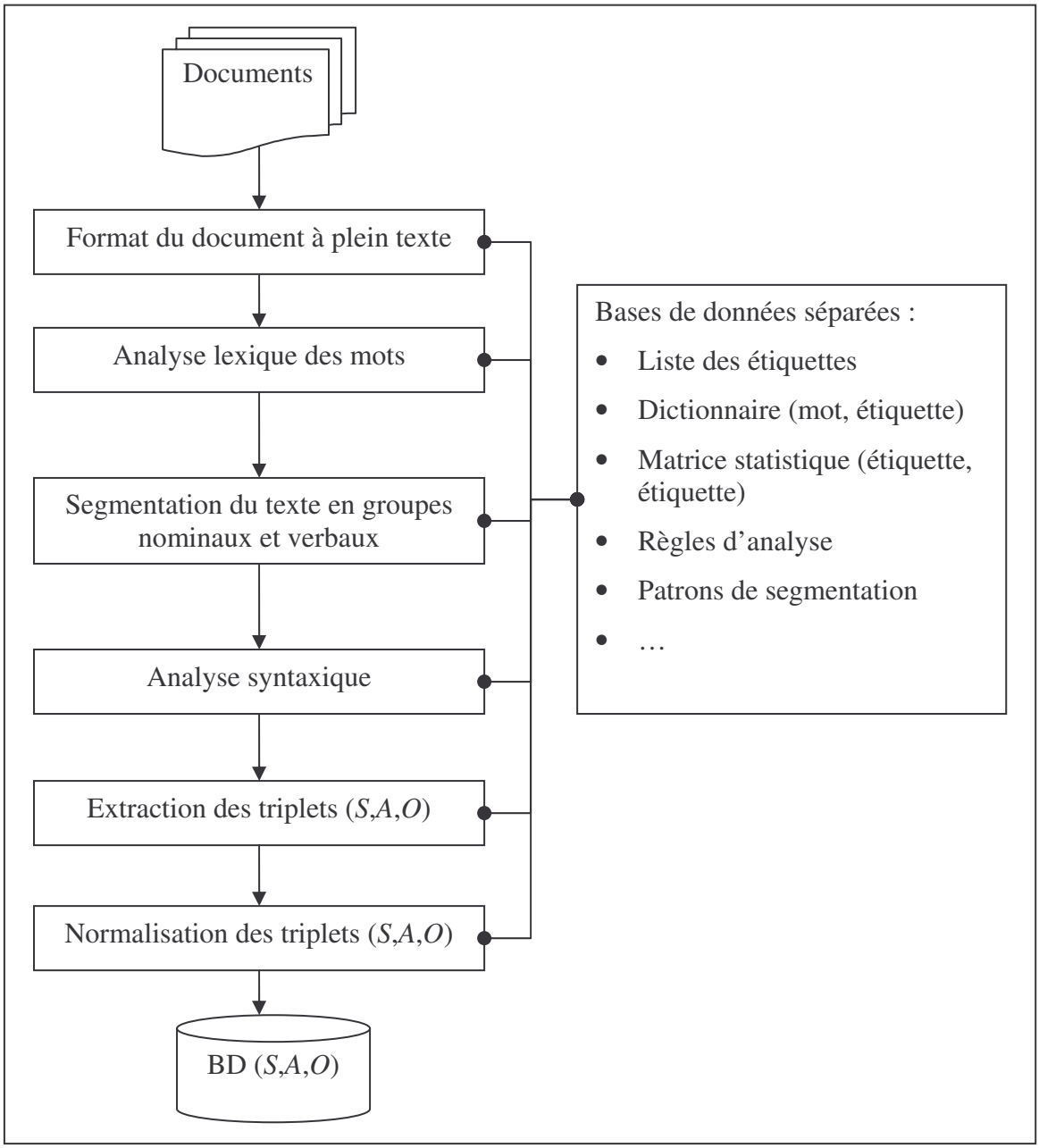


Figure 3.8: Les étapes de traitement nécessaires pour l'extraction des triplets (*S*,*A*,*O*)

1. La difficulté pour repérer les objets et les associer à leur verbe. Dans une phrase, on peut trouver plusieurs sujets, verbes ou objets imbriqués. Pour extraire les triplets, des règles d'extraction sont indispensables (cf. annexe E).

[8] Une action représente toutes les formes d'un verbe avec ses nominalisations

2. En général, les analyseurs syntaxiques ne reconnaissent pas les phrases complexes contenant des clauses imbriquées ou une conjonction de clauses. Par conséquent, l'extraction des triplets devient impossible (cf. annexe E).

3. Comme présenté dans la section 3.4.2, une ressource peut jouer un rôle dans un opérateur différent des rôles de termes associés dans les textes. L'extraction peut ainsi être complètement perturbée.

4. L'analyse syntaxique permet au mieux de repérer des relations entre un verbe, son sujet et son objet. Les nominalisations des verbes sont aussi des indicateurs pour le repérage des opérateurs dans les textes. La relation entre une nominalisation d'un verbe et son sujet ou son objet ne peut pas être extraite par cette analyse syntaxique. Par exemple, dans la phrase "*An alternating magnetic field is applied parallel to the axis of a coil*", le mot *alternating* représente une action dont le sujet est *magnetic field*. L'analyse syntaxique le considère comme un adjectif.

3.6 Bilan

Dans ce chapitre nous avons présenté l'architecture que nous proposons pour un système d'aide à l'innovation, ses modules et leurs tâches. Le module principal de ce système est un sous-système d'extraction dont la tâche est d'extraire des exemples de résolution innovante des problèmes dans un domaine de connaissances. Au contraire des systèmes d'extraction présentés dans le chapitre 2, la tâche de ce module est de localiser des entités sémantiquement définies dans le domaine d'application : ce sont les ressources et les opérateurs d'innovation. L'acquisition manuelle et automatique des règles d'extraction pour ces types d'entités nous est apparue très difficile. En effet l'extraction de ces entités ne peut pas être faite seulement à partir d'indicateurs syntaxiques. Nous avons donc mis en place une base de connaissances dans laquelle ces entités sont sémantiquement représentées pour guider l'élaboration des règles d'extraction.

Chapitre 4

Présentation détaillée de l'approche proposée

Puisque les approches d'extraction d'information fondées sur le traitement automatique des langues naturelles (TALN) sont peu performantes et complexes à mettre en œuvre, nous proposons dans ce chapitre une approche d'extraction par la désambiguïsation sémantique des sens des termes dans les textes. Notre approche de désambiguïsation se base sur une représentation sémantique des opérateurs d'innovation et sur un mapping *Map* de leurs ressources à des ensembles de termes en langue naturelle LN. Une ontologie d'innovation O_{ino} est proposée pour permettre la représentation sémantique des opérateurs d'innovation dans un domaine de connaissances.

4.1 Choix de UML pour la représentation de l'ontologie

De nombreux langages ont été proposés pour décrire des connaissances. Les plus connus et les plus utilisés sont les langages de description logique (KIF, Cycle, OWL, logique premier ordre,…). UML (*Unified Modeling Language*) peut aussi être utilisé pour décrire des connaissances [113,114].

Les langages de description logique sont développés pour être exploités par un système de raisonnement logique. Les connaissances sont représentées par des formules logiques. Le rôle du système de raisonnement est de comparer logiquement ces formules pour identifier leurs relations et inférer de nouvelles connaissances. L'absence d'une connaissance ne veut pas dire qu'elle est fausse. Elle est vraie ou fausse si on peut le prouver logiquement (raisonnement dans un monde ouvert [115,116]). Les formules logiques permettent de représenter les connaissances indépendamment de leur structuration dans les ressources. Une représentation logique est bien adaptée pour une représentation uniforme des ressources de connaissances provenant de sources différentes. Pour l'extraction d'information de textes en langue naturelle, une représentation logique est difficile et peu appropriée.

En revanche, une représentation UML est plus facile à maîtriser et à mettre en œuvre. De plus un système construit autour d'UML peut facilement être intégré et interconnecté à d'autres systèmes. UML permet une description graphique des concepts et des relations. Deux types de diagrammes UML sont plus particulièrement intéressants pour notre approche.

1. Le diagramme de classes permet de modéliser des éléments statiques comme les classes, les attributs et les relations. Ce diagramme est utilisé pour représenter les concepts de notre ontologie (cf. § 4.2).

2. Le diagramme d'objets peut être interprété comme une représentation déclarative des connaissances (cf. § 4.3). Ce diagramme est utilisé pour représenter des instances de concepts c'est-à-dire les opérateurs d'innovation à introduire dans la base de connaissances BC.

Ainsi, UML peut être utilisé pour représenter les entités sémantiques du domaine de connaissances par une ontologie. Le problème d'UML est qu'il est seulement un langage de conception. La représentation graphique n'est pas utilisable pour le raisonnement. Pour supporter le raisonnement nous proposons de convertir le schéma UML sous forme RDF/RDFS. Le raisonnement et les traitements spécifiques à notre approche seront implantés par des méthodes associées aux classes.

Nous illustrons ci-après le principe d'une extraction basée sur une analyse superficielle du texte (cf. diagramme de la figure 4.1). Dans ce diagramme, un concept *Operator* est défini et est associé à deux autre concepts *Action* et *Resource*. Ce concept est prévu pour représenter des connaissances à extraire à partir des textes en langue naturelle. Ainsi, dans les classes correspondant à ces concepts, on a spécifié des opérations (*getTerm* et *getExample*) pour permettre cette extraction par un traitement sélectif et superficiel des textes :

L'opération *getTerm* définie sur la classe *Resource* extrait à partir d'une phrase *ph* les termes correspondant aux instances de cette classe définies dans la base de connaissances. L'attribut *terms* de cette classe mémorise les termes correspondant à chaque instance (i.e. *artefact*, *charge* de la figure 4.2).

L'opération *getExample* de la classe *Operator* vérifie ensuite dans *ph* qu'un terme qui représente une instance de la classe *Resource* est un objet d'un terme qui représente une

instance de la classe *Action*. Elle utilise ces termes pour représenter l'exemple par une instance de la classe *Example*.

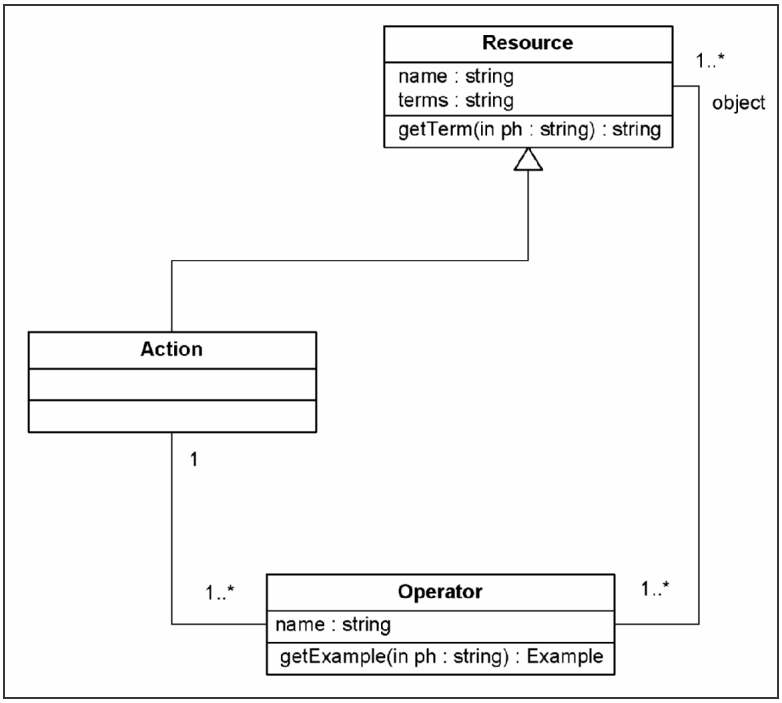


Figure 4.1: Exemple des entités sémantique associé par des relations syntaxiques

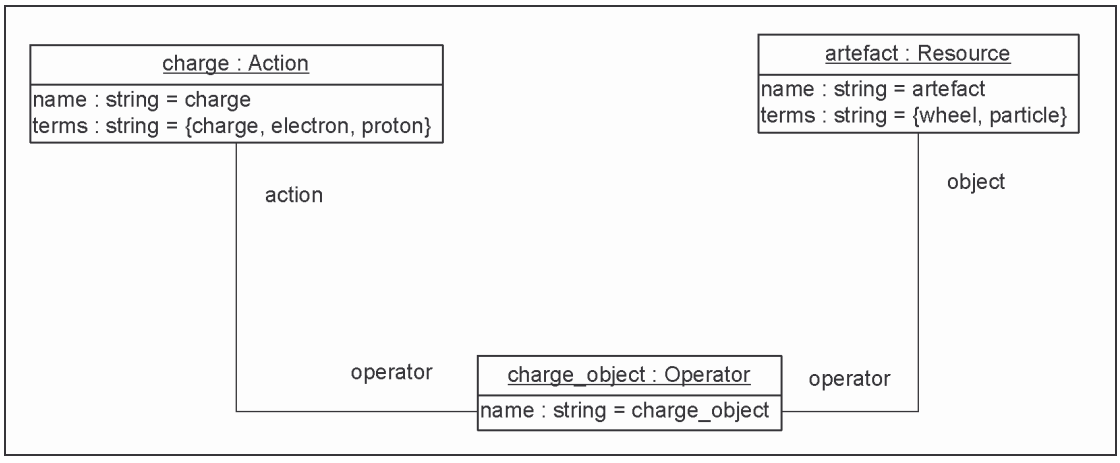


Figure 4.2: Exemple d'une information sémantiquement extraite d'un texte en langue naturelle

Dans la figure 4.2, on a défini une instance *charge_object* du concept *Operator*. Cette instance est associée à une instance *charge* du concept *Action* et à une instance *artefact* du concept *Resource*. A l'instance *charge*, les termes "*charge*", "*electron*", "*proton*" sont associés par son attribut *terms*. Les termes "*wheel*", "*particle*" sont aussi associés à l'instance

artefact. Par exemple, à partir de la phrase *ph* = "*a charge is applied to the abrasive wheel*", l'opération *getExample* de l'instance *charge_object* doit extraire comme exemple le terme "*wheel*" avec le terme "*charge*" comme objet.

4.2 Ontologie d'innovation

Une ontologie est une spécification d'une conceptualisation d'un domaine d'application. Cette spécification définit tous les concepts et les relations nécessaires pour une application. Elle permet à des domaines de connaissances partageant cette ontologie de partager aussi l'application. Une des applications les plus répandues des ontologies est l'intégration des sources de données. Des sources de données présentent leur contenu par rapport à une ontologie donnée. Il est alors possible de les exploiter à travers un système de médiation pour répondre à des requêtes posées par des utilisateurs ou par des agents.

Notre ontologie d'innovation O_{ino} est représentée par trois diagrammes de classes UML dans les figures 4.3 à 4.5 Ces diagrammes représentent respectivement les trois concepts centraux (*Resource*, *Operator* et *Example*), leurs relations et leurs sous-concepts.

Dans ces diagrammes, les concepts sont représentés par des classes. Leurs propriétés sont représentées par des attributs (i.e. *name* des concepts *Resource* et *Operator*) et par des associations (i.e. *improve* entre les concepts *Resource* et *Operator*).

En UML, une classe est un rectangle de trois cases. La première case spécifie le nom de la classe. La deuxième case spécifie les attributs et la troisième case spécifie les opérations effectuant un traitement local sur les instances de la classe (objets).

Pour une implémentation orientée objet, UML définit plusieurs types de relations entre classes. Ces relations peuvent être utilisées pour représenter sémantiquement des rôles et des relations définis dans un domaine de connaissances :

- La relation de généralisation/spécialisation entre concepts (qui correspond à une relation d'héritage) est utilisée pour définir les types des opérateurs d'innovation et les types des ressources.

- Les relations sémantiques partie/tout entre concepts peuvent être représentées par la relation de composition entre classes. Elle signifie que deux objets étant différents, leurs parties doivent être différentes. Cette relation est représentée en UML par une ligne ayant un diamant noir à son extrémité (cf. *Artefact* de la figure 4.3).

4.2.1 Concept *Resource*

Le concept *Ressource* (cf. figure 4.3) représente tout ce qu'on peut utiliser pour résoudre ou améliorer un problème. Ses sous-concepts représentent les types des ressources présentées dans le chapitre 1 (cf. § 1.2.3). Les hiérarchies associées sont représentées par une relation de généralisation/spécialisation (Hyperonymie/Hoponymie). Trois types de ressources sont utilisés pour représenter les opérateurs d'innovation (cf. figure 4.4). Ils sont définis par les concepts *Resource*, *State* et *Parameter*. Les autres types de ressources comme l'énergie et les objets naturels utilisés dans l'innovation sont respectivement représentés par les classes Energy et Physical_object.

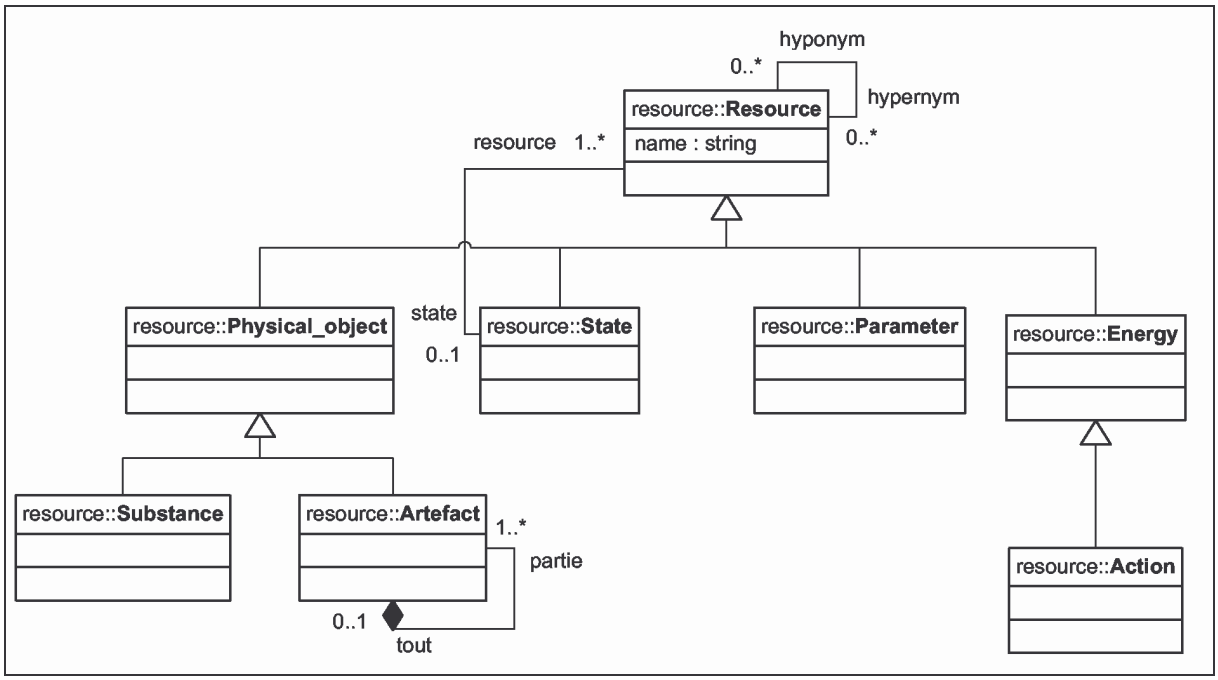


Figure 4.3: Hiérarchies des ressources

Dans notre organisation des ressources, on note que les actions sont définies par un sous-concept de *Energy*. Cette spécialisation concerne les verbes qui correspondent à des actions de transformation d'énergie (i.e. *segmentation*, *charge*, *alternate*, *reflect*) par opposition aux

autres types de verbes (i.e. *be*, *have*, *can*, *consiste of*) qui introduisent une description d'entités.

Les objets physiques, représentés par le concept *Physical_object*, sont divisés en deux hiérarchies *Artefact* et *Substance*. Une instance du concept *Artefact* peut être composée d'autres artéfacts. La relation de composition sur la classe permet d'associer chaque instance à ses composants.

Le concept *State* représente l'état normal d'une ressource. Une ressource ne peut avoir qu'un seul état à un instant donné (*solide*, *liquide*, *gaz*, *poudre*, *plasma*). Cet état est défini par une association entre les classes *State* et *Resource* respectivement. Par conséquent, le changement d'état d'une ressource par un opérateur doit changer la ressource. Cette information est nécessaire pour représenter les solutions innovantes génériques (cf. § 4.3).

Le concept *Parameter* représente les 39 paramètres d'innovation définis dans la méthode TRIZ (cf. § 1.2.12). Ces paramètres sont mis en relation de contradiction par les principes inventifs (cf. § 1.2.13.1).

4.2.2 Concept *Operator*

Les opérateurs d'innovation sont des recommandations (cf. § 1.2.13) à utiliser pour résoudre d'une manière inventive les problèmes détectés dans les systèmes techniques et pour les améliorer. Dans le diagramme 4.4, le concept *Operator* représente les différents opérateurs et le problème qu'ils améliorent est défini par les ressources associées par le rôle *improve*. Les trois types des opérateurs définis dans la méthode TRIZ (effet, solution standard et principe) sont représentés par les classes *Effect*, *Principle* et *Standard* respectivement. Un effet est une relation cause-conséquence. La cause est définie par des ressources associées par le rôle *cause* à la classe *Effect*. La conséquence est le problème amélioré et est définie par le rôle *improve* du concept père *Operator*.

Un principe se base sur des paramètres en contradiction pour résoudre le problème associé. La relation entre le concept *Principle* et le concept *Paramètre* permet d'exprimer cette contradiction entre paramètres pour chaque principe. Autrement dit, dans chaque représentation d'un principe inventif, les paramètres associés sont supposés tous contradictoires par rapport à cette représentation (cf. § 4.3).

Une solution innovante générique (cf. § 1.2.13.4) résout un problème en améliorant la liaison entre ses composants. Les états de ces composants sont les ressources à utiliser pour résoudre le problème. Dans le digramme 4.4, les solutions et les états sont représentés respectivement par les classes *State* et *Standard*. L'association entre ces classes permet de définir les états correspondant à chaque solution.

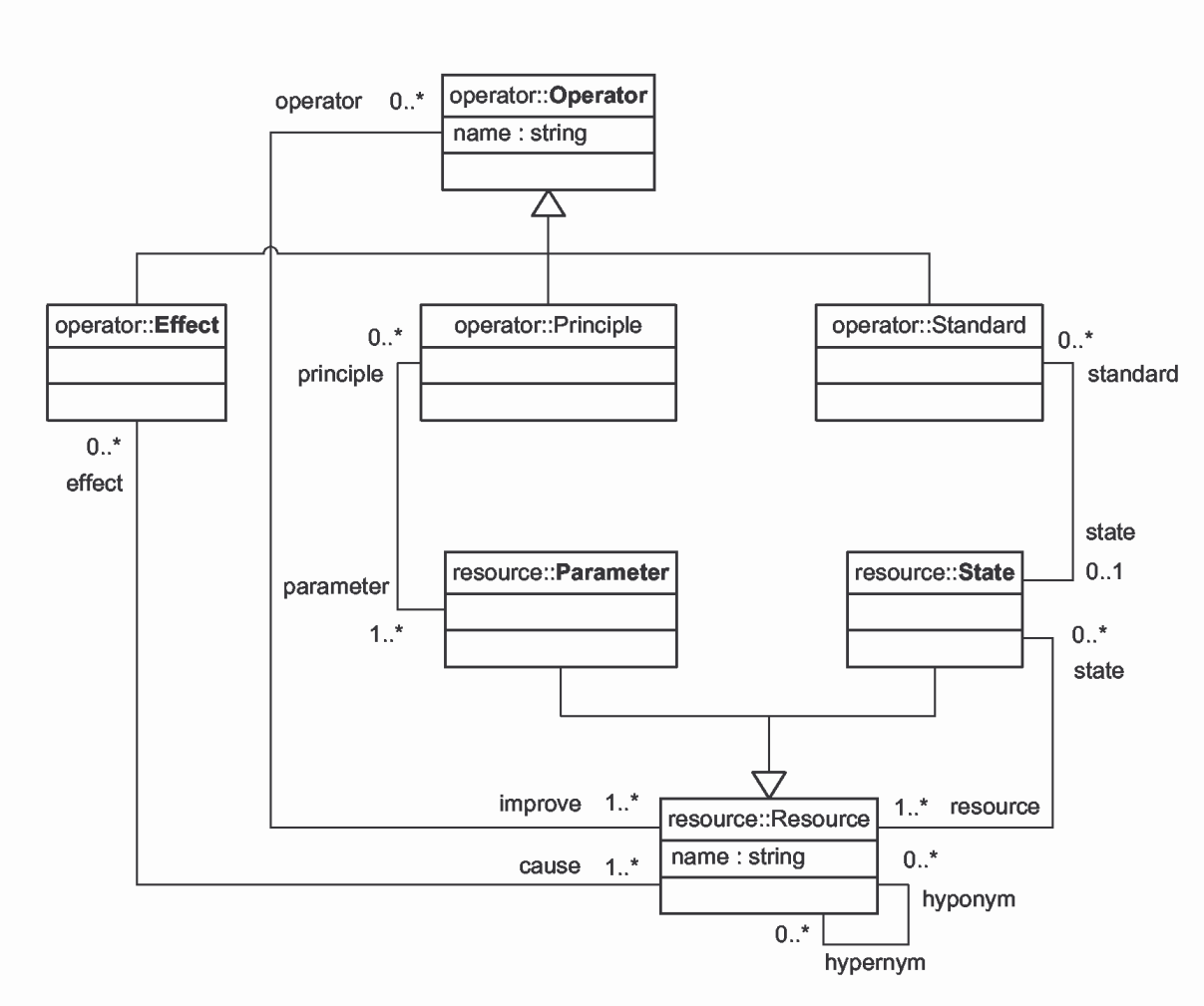


Figure 4.4: Concept *Operator*

4.2.3 Concept *Example*

Le concept *Example* représente sémantiquement les exemples illustrant l'utilisation des opérateurs. Chaque instance de la classe *Example* représente un opérateur dans le texte en langue naturelle. Cet opérateur doit être associé à l'exemple. L'attribut *text_link* est le lien vers ce texte. Chaque instance de *Example* associe des termes repérés à leurs ressources dans la base de connaissances. Ces termes sont des instances de la classe *Term.* Un texte peut

illustrer plusieurs exemples d'opérateurs. Le concept *Example* représente le formulaire à extraire pour les opérateurs d'innovation. Ainsi on peut l'utiliser pour définir automatiquement des formulaires d'extraction. Une approche d'extraction est proposée dans ce chapitre (cf. § 4.5) pour extraire les exemples et les représenter sémantiquement dans la base de connaissances BC. Dans le chapitre suivant, nous associons des informations syntaxiques et sémantiques définies par le mapping *Map* et nous les utilisons pour implémenter une base de données relationnelle pour l'extraction.

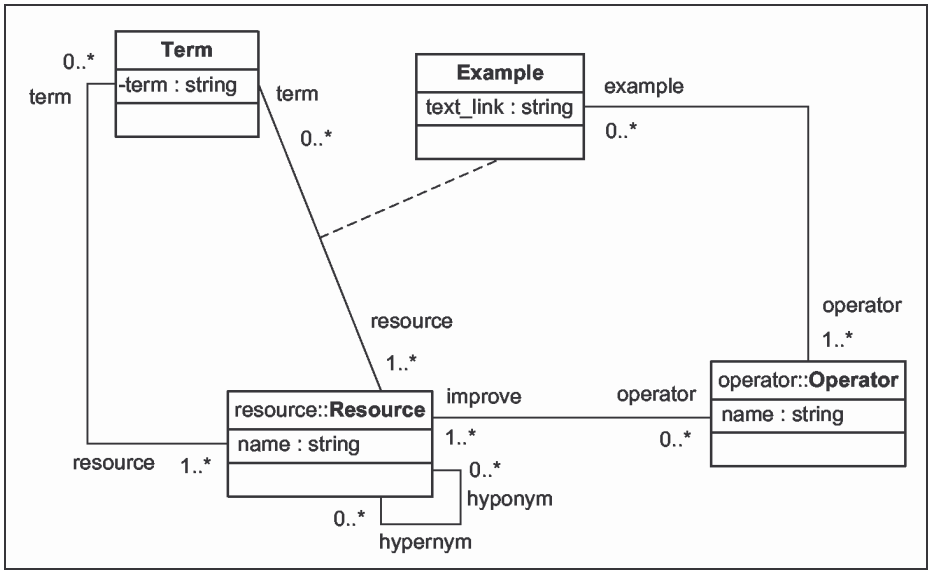


Figure 4.5: Concept *Example*

4.3 Base de connaissances BC

Notre ontologie O_{ino} définit les concepts et leurs relations nécessaires pour l'extraction des exemples des opérateurs d'innovation d'un domaine de connaissances. Par cette ontologie un expert peut représenter les opérateurs et les ressources comme des instances de ses concepts. Pour chaque concept (opérateur ou ressource) défini dans l'ontologie, l'expert peut introduire dans une base de connaissances BC une ou plusieurs instances. L'attribut *name* des classes *Ressource* et *Operator* (cf. figure 4.3 et 4.4) est ajouté pour permettre le repérage de leurs instances et l'annotation sémantique des textes.

Dans les figures 4.6, 4.7 et 4.8, on a défini en UML par des diagrammes d'objets une représentation sémantique de chacun des trois types d'opérateurs (effet, principe et solution standard).

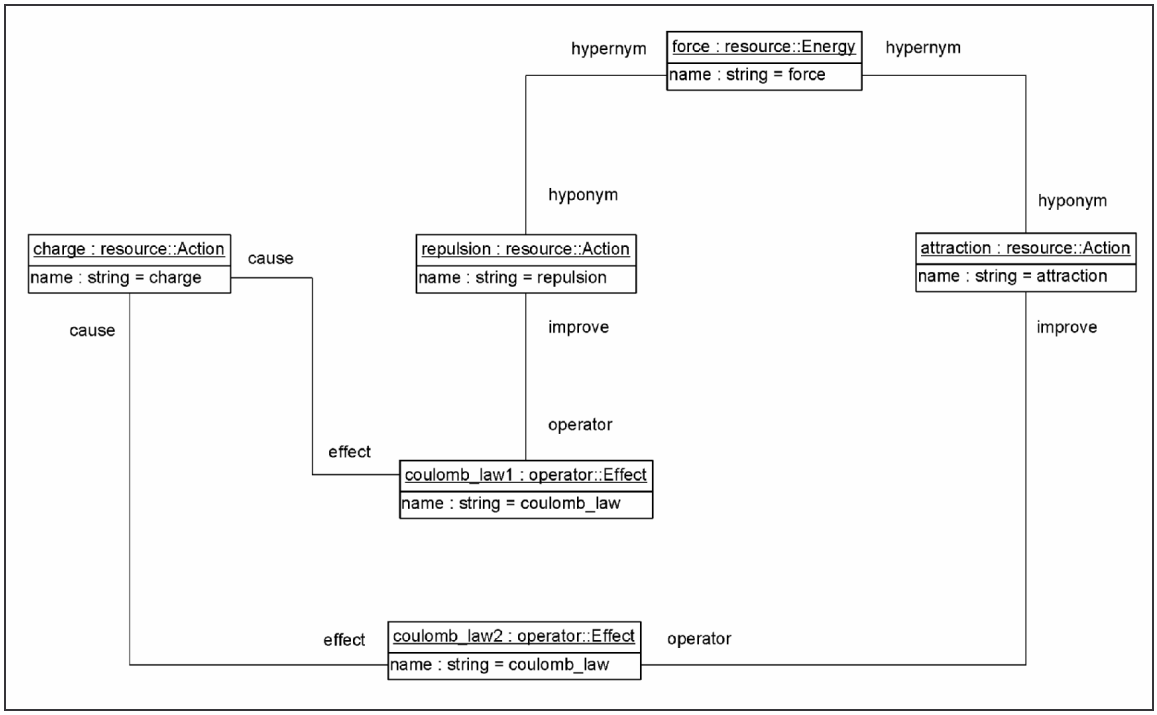


Figure 4.6: Représentation sémantique de l'effet loi de Coulomb

La figure 4.6 représente l'effet loi de Coulomb qui est défini par deux instances c*oulomb_law1*, c*oulomb_law*2 de la classe *Effect.* Dans l'instance *coulomb_law1*, l'effet est défini par deux instances *charge* et *repulsion* de la classe *Action*. Associée par le rôle *cause*, l'instance *charge* représente la cause de l'effet. L'instance *repulsion* qui est associée par le rôle *improve* représente la conséquence. Puisque cet opérateur peut avoir une autre conséquence qui est la force d'attraction représentée par *attraction*, l'instance c*oulomb_law2* est introduite pour la définir. Dans la représentation de cet effet, les actions *repulsion* et *attraction* sont spécifiées comme des forces en les associant (par le rôle hyperonyme) à une instance *force* du concept *Resource*.

Représentée par l'instance *segmentation* de la classe *Action*, la segmention est la ressource à améliorer par les opérateurs présentés dans les deux figures 4.7 et 4.8. Dans la figure 4.7, l'opérateur défini est le principe de segmentation dont le rôle est d'améliorer la segmentation par la mise en contradiction des paramètres d'innovation. Deux instances *segmentation5* et *segmentation6* de la classe *Principle* sont définies à partir de la matrice de contradiction (cf. § 1.2.13.1). La première instance *segmentation5* met en contradiction les paramètres forme et facilité de réparation. L'instance *segmentation6* met en contradiction la surface d'un objet fixe et la pression. Les quatre paramètres utilisés forme, facilité de réparation, surface d'un objet fixe et pression sont représentés respectivement dans le

diagramme d'objets par les instances *shape*, *repairability*, *area_of_nonmoving_object*, *pressure* de la classe *Parameter* (cf. figure 4.3).

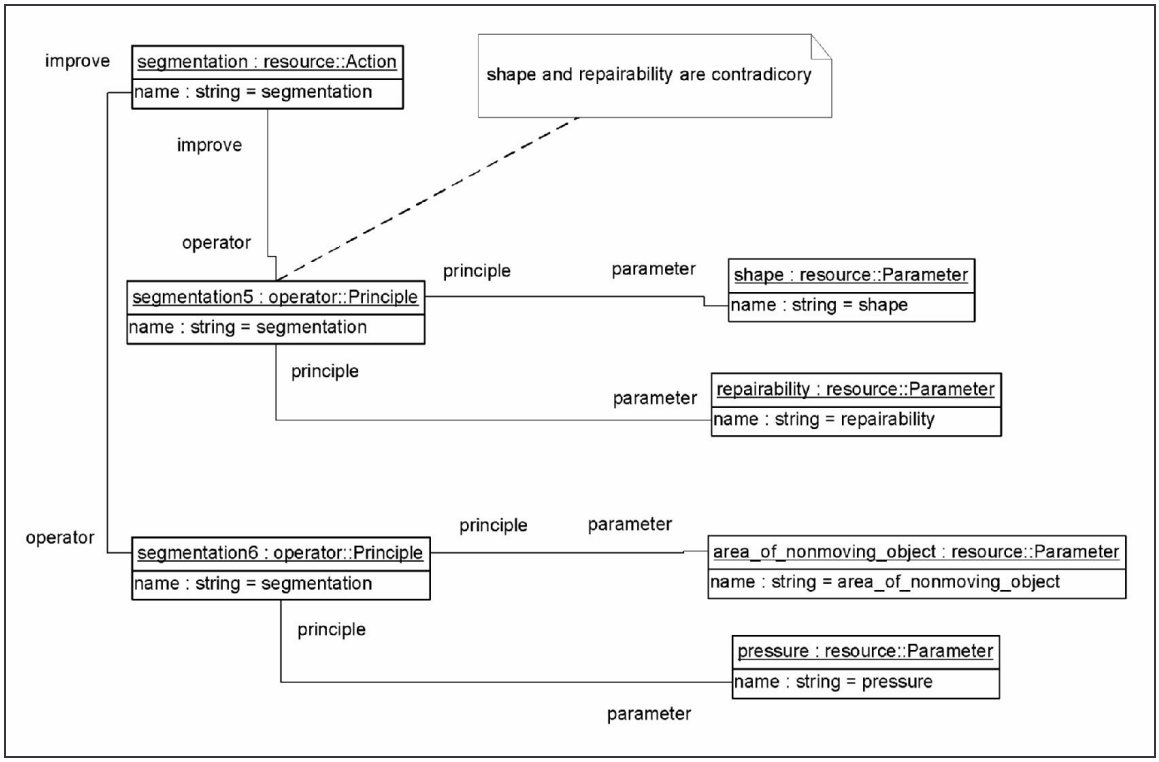


Figure 4.7: Représentation sémantique d'un principe de segmentation mettant en contradiction deux attributs, la forme avec la stabilité

Dans la figure 4.8, c'est une solution innovante générique qui est suggérée pour améliorer la segmentation à partir des différents états possibles (solide, liquide, gaz et poudre). Chaque état représente une solution pour effectuer la segmentation. Ainsi quatre instances (*segmentation1*, *segmentation2*, *segmentation3*, *segmentation4*) de la classe *Standard* sont définies pour représenter respectivement les solutions correspondantes aux états précédents. Ces états sont représentés respectivement par les instances (*solid*, *liquid*, *gas*, *powder*) de la classe *State* et sont associés aux solutions correspondantes.

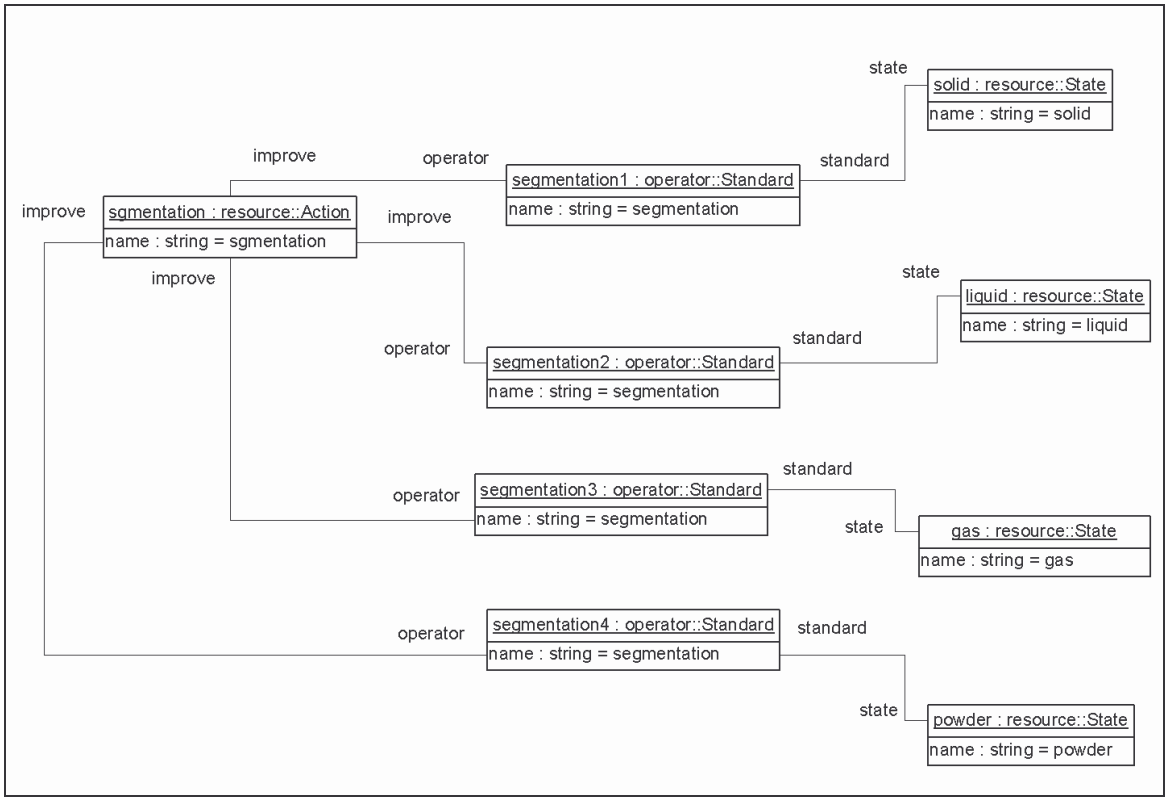


Figure 4.8: Représentation sémantique d'une solution inventive pour améliorer une segmentation

4.4 Représentation des exemples

Un système de recherche et d'extraction (cf. chapitre 3) est conçu pour extraire les exemples à partir des textes en langue naturelle. Chaque instance du concept *Example* est extraite par un formulaire d'extraction. Dans la figure 4.9, nous présentons un exemple pour l'opérateur loi de Coulomb à partir du formulaire d'extraction donné dans la figure 3.4. Dans cette représentation, nous avons associé l'exemple à ses différentes ressources. Le lien entre le texte et l'exemple est représenté par un attribut (non représenté sur le diagramme) de la classe *Example*.

On distingue dans la figure 4.9 les ressources propres à l'opérateur loi de Coulomb et les ressources objets associées:

1. Les ressources (*charge*, *repulsion*) utilisées pour représenter l'opérateur de la figure 4.6. sont les ressources spécifiques. Les termes extraits du texte sont "*charge*", "*repel*".

2. Les autres ressources sont les ressources objets. Dans l'exemple, une ressource objet est repérée. Elle est de type *physical_object*. Elle est associée à deux termes du texte : "*surface of wheel*" et "*material particle*".

Dans la représentation sémantique, on a évité d'utiliser des noms de rôle (sujet, action, objet) pour deux raisons :

1. Il ne faut pas les confondre avec les rôles syntaxiques qui ne leur correspondent pas nécessairement (cf. chapitre 3).

2. La représentation sémantique des opérateurs associés à l'exemple est suffisante pour les distinguer.

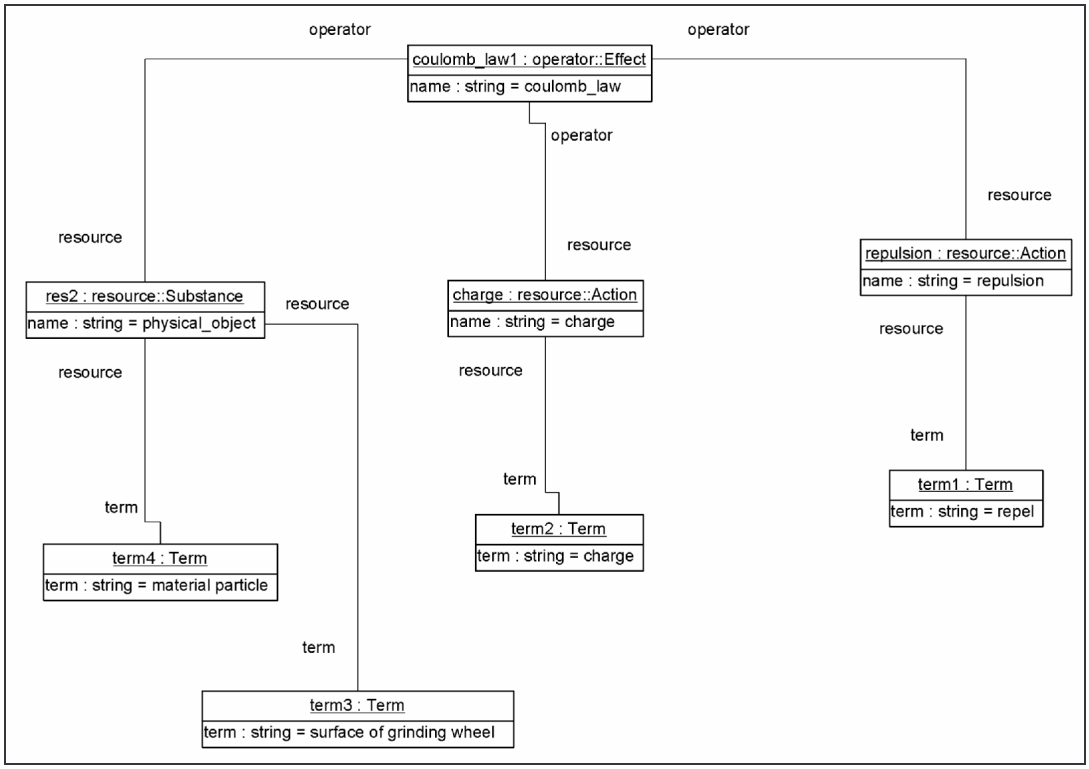


Figure 4.9: Représentation sémantique de l'exemple de l'effet loi de Coulomb

4.5 Extraction des exemples à partir des textes en langue naturelle

Pour l'extraction des exemples à partir des textes en langue naturelle, il faut définir des formulaires d'extractions. Les opérateurs d'innovation représentés dans la base de connaissances BC définissent les informations à extraire des textes. Le concept *Example* de

notre ontologie peut être utilisé pour définir le formulaire à extraire pour chaque opérateur. En effet sont associées à ce concept toutes les informations nécessaires sauf les ressources sur lesquelles un opérateur est réalisé (cf. § 4.3). Ces ressources ne sont pas connues à l'avance. Pour les repérer et les extraire, nous utiliserons les relations syntaxiques avec les ressources de l'opérateur.

Dans notre approche d'extraction, nous avons exclu de définir des règles d'extraction linguistique, à cause de la difficulté de trouver des règles pertinentes et générales. L'approche adoptée se base essentiellement sur une désambiguïsation du sens des termes dans les textes.

La désambiguïsation du sens des termes dans les textes en langue naturelle se base sur des ressources de connaissances linguistiques et sémantiques (dictionnaires électroniques, thesaurus, réseaux sémantiques, ontologies) [117]. Les termes apparus dans le contexte de chaque terme *t* dans le texte sont utilisés pour identifier les ressources qui correspondent le mieux à *t*. Ainsi, la clé de cette technique ne réside pas dans des patrons linguistiques mais dans les relations sémantiques des termes dans la base de connaissances. Dans certains travaux où des ressources générales de connaissances sont utilisées (i.e. WordNet), des fonctions de similarité peuvent aussi être spécifiées pour réaliser la désambiguïsation (cf. § 2.1.3.2).

Notre approche de désambiguïsation présentée ci-après, s'appuie sur la base de connaissances BC et sur un mapping *Map* (cf. § 4.6) des ressources à des termes candidats dans une langue naturelle LN. Par cette approche, on ne fait pas une désambiguïsation totale de texte. On se focalise seulement sur des mots candidat (défini par le mapping *Map*) qui peut représenter des ressources. Le principe de l'approche est décrit ci-après.

Pour chaque terme *t* dans un texte candidat *x*:

1. Identifier les ressources mappées à ce terme. Si aucune ressource n'est identifiable, le terme doit être rejeté.

2. Identifier tous les opérateurs de la base BC dont les ressources sont représentées par des termes candidats dans le contexte de *t* et différents de *t*.

3. Pour chaque opérateur *p*, le texte *x* est un exemple pour *p*. Le terme *t* représente une ressource objet sur laquelle *p* est appliqué. Le terme *t* avec les termes qui représentent les ressources de *p* représentent des ressources de *x*.

4. Pour associer *t* aux ressources les plus appropriées dans la base de connaissances, on doit repérer un opérateur dans la base de connaissances dont *t* est une ressource spécifique (on doit utiliser aussi la relation d'hyperonymie entre ressources).

Dans les approches de désambiguïsation [117, 118, 119, 120], le contexte est défini par un ensemble de mots avant et après le mot *t* dans le texte indépendamment des phrases. Dans notre approche, on définit le contexte par les phrases où apparaît le terme candidat *t*. Ainsi, un terme *t'* appartient au contexte de *t*, si *t* et *t'* apparaissent tous les deux dans une même phrase. Nous n'avons utilisé aucune fonction de similarité entre termes, car on suppose que la base de connaissances, ses opérateurs et leurs relations constituent des informations suffisantes pour permettre la désambiguïsation. Cette hypothèse se base sur l'idée que la base de connaissances contient toutes les informations nécessaires (les ressources, les opérateurs et leurs relations sémantique) pour la désambiguïsation par l'approche proposée. Notre mapping *Map* n'a qu'un but, c'est d'associer hiérarchiquement les termes à leurs ressources pour permettre leur repérage sémantique dans les textes.

Pour illustrer cette approche, reprenons le texte présenté dans la figure 3.4. Ce texte est un exemple de représentation sémantique de l'instance *coulomb_law1* (cf. figure 4.6) de l'opérateur loi de Coulomb. Les ressources spécifiques de cette représentation sont *charge* et *repulsion.* Supposons que le mapping *Map* défini par l'expert associe le terme "*charge*" à la ressource *charge* et le terme "*repel*" à la ressource *repulsion*. Et soient "*potential*", "*voltage*", "*machine*", "*process*", "*wheel*", "*time*", "*piece*", "*particle*", "*material*", "*surface*" des termes ambigus représentent des ressources candidates dans la base BC par le mapping *Map*.

Parmi ces termes candidats, les termes pertinents sont "*particle*", "*machine*", "*material*", "*surface*", "*wheel*", car chacun d'eux apparaît dans une phrase avec les termes "*charge*" et "*repel*" mappés aux ressources *charge* et *repulsion* de *coulomb_law1*. Par conséquent, ces termes représentent les ressources objet de l'opérateur. Ces termes peuvent être associés à plusieurs ressources dans la base de connaissances. Pour repérer les ressources les plus pertinentes, en prend en compte les ressources associés à d'autres exemples dans le même texte. L'association de ces termes à leurs exemples nous fournit toutes les informations

sémantiques nécessaires pour l'annotation automatique du texte (i.e. terme, ressource, opérateur). Pour annoter le texte (cf. figure 3.2), on peut effectuer une segmentation en groupes nominaux et verbaux de la phrase contenant les termes extraits. Chaque groupe contenant des termes extraits est encadré par les balises <res> et <res/> où res est le nom de la ressource associée.

Dans nos expérimentations (cf. chapitre 5), pour simplifier cette approche, le prototype implémenté effectue la première étape sur la totalité d'un ou plusieurs textes. Des tables relationnelles sont conçues pour stocker les annotations et le résultat des traitements effectués sur le texte. Cette manipulation facilite le repérage répétitif des opérateurs apparus dans le contexte de chaque terme candidat *t*. En nous basant sur l'ontologie O_{ino}, des requêtes SQL sont composées pour permettre la désambiguïsation par des tables relationnelles représentant des informations superficielles extraites du texte.

4.6 Mapping des ressources à des termes

Un mapping $F : G_1(V_1, E_1) \rightarrow G_2(V_2, E_2)$ est un morphisme i.e. une fonction telle que :

$\forall e_1 = vu \in E_1, F(e_1) = F(v)F(u) = e_2$ et $e_2 \in E_2$ où :

$G_i(V_i, E_i)$ et i=1,2 est un graphe, V_i l'ensemble de ses sommets et E_i est un ensemble d'arêtes entre ces sommets.

Selon l'ontologie d'innovation O_{ino}, les ressources (cf. figure 4.3) sont représentées selon des hiérarchies organisées par la relation d'hyperonymie/hyponymie. Cette relation qu'on suppose réflexive, transitive et antisymétrique définit un ordre partiel sur l'ensemble *S* des ressources de la base de connaissances. Cet ordre peut être représenté par un graphe $(S, \leq_h)$ où :

- *S* représente les sommets du graphe.
- $(\leq_h)$ définit des arêtes entre les sommets.
- $\forall s, s' \in (S, \leq_h), s \leq_h s' \Leftrightarrow s'$ est hyperonyme de s.

Dans les textes en langue naturelle, les ressources de S ne sont identifiables que par des ensembles de termes associés à ces ressources. Les termes qui identifient sémantiquement une ressource, identifient aussi toutes ses ressources hyperonymes. Soit T l'ensemble de tous les termes du texte à associer aux ressources de l'ensemble S des ressources. Et soit $(2^T, \subseteq)$ un graphe dans lequel les sous-ensembles de T sont partiellement ordonnés par la relation d'inclusion ($\subseteq$). Ainsi, un mapping *Map* du graphe $(S, \leq_h)$ au graphe $(2^T, \subseteq)$ permet :

1. de calculer à partir d'un terme t dans un texte x l'ensemble S_t des ressources candidates associées à t par le mapping *Map* où $S_t = \bigcup_{s \in S \wedge t \in Map(s)} \{s\}$;

2. de minimiser le travail de l'expert pour définir le mapping. En effet, dès que l'expert mappe une ressource s à un ensemble de termes $T' = Map(s) \subseteq T$, ces termes sont alors automatiquement associés par *Map* à toutes les ressources hyperonymes de s, car le mapping vérifie la propriété suivante:

$$\forall s, s' \in S \wedge s \leq_h s' \Rightarrow Map(s) \subseteq Map(s')$$

4.7 Utilisation de *WordNet* pour effectuer et faciliter le mapping

4.7.1 Présentation de *WordNet*

Développée à l'université de Princeton, WordNet [110] est une base de données lexicale de la langue anglaise. Dans cette base, les mots anglais sont associés par deux types de relations : des relations lexicales et des relations sémantiques. Par conséquent cette base forme un réseau sémantique. Dans la figure 4.10, nous présentons en UML le diagramme de classes avec les concepts et les relations intéressants pour notre travail.

Le concept *Word* représente les mots de la langue. Chaque instance de ce concept représente un mot défini par l'attribut *word*. Les mots sont associés par des relations lexicales comme la relation d'antonymie (*antonym*). C'est une relation de contradiction (i.e. "*small*" est antonyme de "*big*"). Une autre relation lexicale est la relation *related_form* qui associe une famille des mots dérivés d'une même racine (i.e. "*machine*", "*machinery*", "*machiniste*").

Les différents sens d'un mot sont représentés par des *synsets* (cf. figure 4.11). Ce sont des instances du concept *Synset.* Chaque *synset* est associé à des instances du concept *Word.* Les instances de *Word* associées à un même *synset* représentent des mots synonymes. Les relations entre *synsets* sont les relations sémantiques dans WordNet. La relation principale est la relation de hyperonymie/hyponymie. Chaque *synset* est défini par un offset, c'est sa clé qui l'identifie dans WordNet. Un *synset* est défini aussi par un glossaire qui l'explique et l'associe à des exemples.

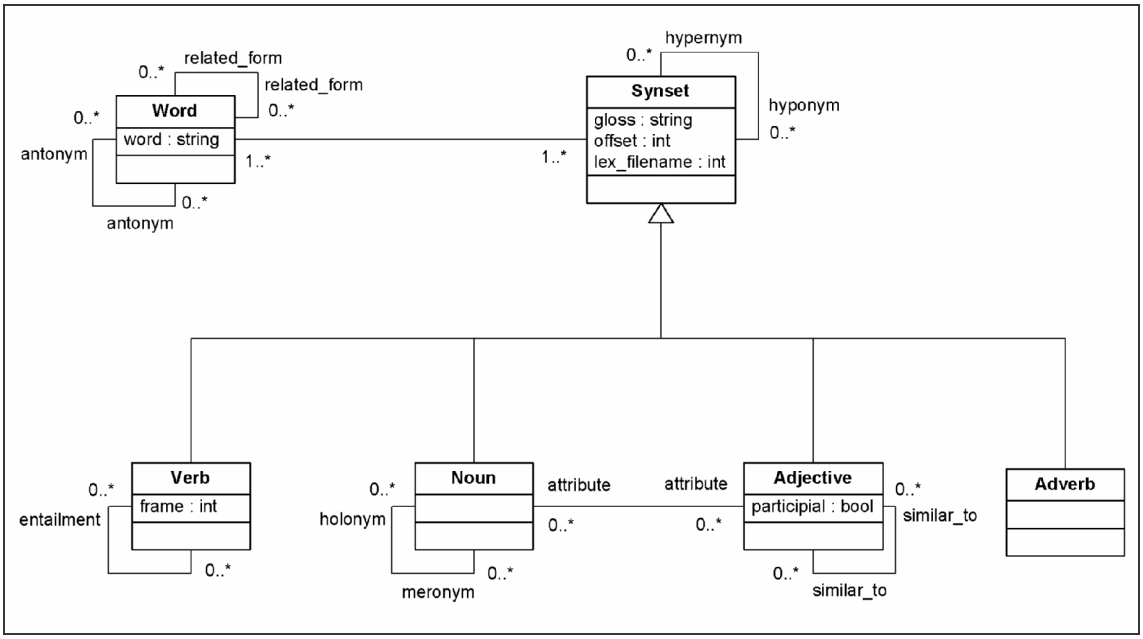


Figure 4.10: Schéma conceptuel (partiel) du réseau sémantique WordNet

Dans WordNet les *synsets* sont séparés en quatre catégories syntaxiques qui sont les verbes, les noms, les adjectives et les adverbes. Ces catégories sont représentées dans le diagramme respectivement par les concepts *Verb*, *Noun*, *Adjective*, *Adverb*.

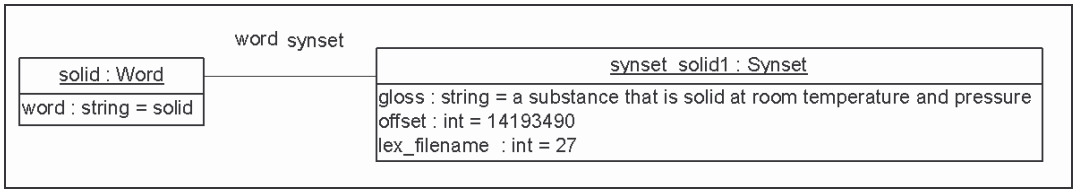


Figure 4.11: Un exemple d'un *synset* de WordNet associé au mot "*solid*"

En plus de la relation d'hyperonymie, les verbes interviennent dans deux autres relations sémantiques. Il s'agit de la relation *entailment* et de la relation *troponym*. La première relation permet d'identifier sémantiquement les verbes qui expriment la même chose qu'un verbe

donné (i.e. *to live* exprime la même chose que *to exist*). La deuxième relation permet d'identifier des verbes qui correspondent à une action similaire à celle d'un verbe donné (i.e. *to decrease* similaire à *to break*, *to taper*, *to flex*, *to drop*,…).

Les noms ont entre eux la relation *meronym/holonym* (partie/tout). Cette relation permet d'associer les *synsets* qui représentent des substances, des artéfacts et des ensembles à leurs parties et à leur membres et vice-versa (i.e. *book* est partie de *text*, *book* a pour partie *cover*).

La relation de synonymie entre adjectifs est exprimée aussi par une relation de similarité (i.e. *solid* synonyme de *sound*). Les adjectifs et les noms sont associés aussi par la relation symétrique *attribute*. Cette association permet d'identifier les noms et leurs adjectifs qui les valorisent (i.e. *big* a attribut *size* et *intensity* a attribut *loud* et *soft*).

Une autre manière de classifier des *synsets* est proposée dans le diagramme UML 4.10 en se basant sur l'attribut *ex_filename* du concept *synset*. Cet attribut sépare les *synsets* de WordNet en quarante-quatre classes lexicographiques (cf. annexe F).

4.7.2 Mapping par le moyen de *WordNet*

L'association de tous les termes possibles par l'expert à chaque ressource définie dans la base de connaissances peut être une tâche très lourde. Pour la faciliter, nous avons pensé à séparer le mapping en deux fonctions de mapping Map_1 et Map_2 où $Map = Map_1 \circ Map_2$ (cf. figure 4.12) où

- $Map(s) = Map_1 \circ Map_2(s) = Map_1(Map_2(s)), \forall s \in S$,
- Map_1 est un mapping du graphe $(S, \leq_h)$ au graphe $(2^N, \subseteq)$.
- Map_2 est un mapping du graphe $(2^N, \subseteq)$ au graphe $(2^T, \subseteq)$.
- S est l'ensemble des ressources dans la base de connaissances BC.
- N est l'ensemble des *synsets* nominaux (*Noun*) dans WordNet.
- T est l'ensemble des termes verbes ou noms existant dans WordNet.

La fonction Map_1 est semi automatique. Un expert mappe une ressource *s* à des *synsets* de *N*. Le système se base sur la définition du mapping pour associer les hyperonymes de *s* à ces *synsets*.

La fonction Map_2 est automatique. Elle se base sur les relations lexicales et sémantiques présentées dans le diagramme 4.12 pour associer les *synsets* aux termes correspondant dans *T*.

Par conséquent selon la définition de composition la fonction *Map* mappe la ressource *s* à ces termes.

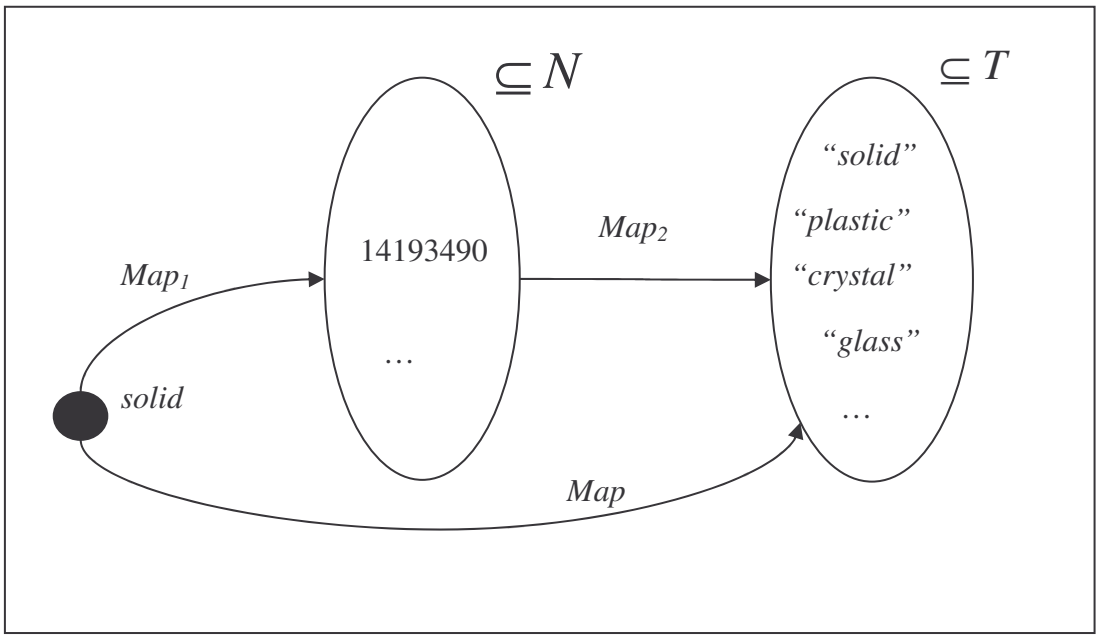


Figure 4.12: Fonction de mapping $Map = Map_1 \circ Map_2$ des ressources aux termes de la langue anglaise

Par ce mapping nous avons résolu aussi des problèmes d'adaptabilité de WordNet à notre besoin :

- Par le mapping *Map*, les relations qui nous intéressent et qui n'existent pas dans WordNet sont extraites à partir de la base de connaissance BC. Par exemple, les termes hyponymes de *solid* dans notre domaine ne sont pas tous associés à ce terme dans WordNet ; on les associe par le mapping *Map* à une ressource *solid* dans notre base de connaissances.

- Par le mapping Map_1, *l*es hiérarchies des *synsets* qui nous intéressent sont regroupées par les hiérarchies des ressources. Donc, les mots inutiles sont automatiquement négligés, car leurs *synsets* ne sont pas associés par le mapping Map_1 aux ressources.

- Les mots qui n'existent pas dans WordNet comme le mot *cavitation* peut être directement ajouté par l'expert aux termes mappés par le mapping *Map* à la ressource correspondante.

4.8 Recherche documentaire sur le Web

Comme présenté dans le chapitre précédent, notre système est conçu pour l'extraction des exemples à partir du Web. Pour le repérage de textes sur le Web un module de recherche documentaire est proposé (cf. § 3.3.1).

Pour le repérage des mots clés nécessaires à la composition automatique des requêtes de recherche sur le Web, nous proposons l'utilisation de l'approche de Santamarìa *et al.* [121] pour repérer des dossiers Web et les associer à des *synsets* dans WordNet.

Leur approche prend en entrée, les *synset* de type *Noun* dans WordNet. Pour un mot *m* et pour chaque sens (*synset*) *n* associé à *m* dans WordNet, le système compose une requête *q* (cf. figure 4.13). La requête *q* est composée d'un mot principal, c'est le mot *m*, de mots positifs optionnels et de mots négatifs. Les mots positifs M^+ sont les mots associés au *synset n* et à ses hyperonymes directs. Les mots négatifs M^- sont les mots qui sont synonymes (cf. § 4.7.1) du mot *m* et qui ne sont par associés au *synset n* (ils sont associés à d'autres *synsets* de *m*). Puis, le système utilise cette requête *q* pour repérer un ensemble *D* des dossiers candidats. Ensuite, le système extrait de WordNet une description plus riche du *synset n* à partir de ses relations (*hypernym*, *hyponym*, *myronim*, *holonym*…). Il la compare avec les dossiers de *D*. Enfin, il l'associe aux dossiers pertinents, en se basant sur le calcul d'un score de confidence.

Dans notre approche, les mots candidats et leur *synsets* de types *Noun* sont définis à partir des ressources par les mappings *Map* et Map_1. Ces fonctions de mapping nous permettent de composer exactement le même type de requêtes de recherche que celui présenté ci-dessus.

Pour chaque *synset n* mappé à une ressource *s*, pour chaque mot *m* associé à ce *synset* dans WordNet, une requête *q* peut être composée. Le mot principal de *q* est *m*, ses mots

positifs M^+optionnels sont les termes mappés par *Map* à *s* et différents de *m* ($M^+ = Map(s)/\{m\}$), ses mots négatifs M^- sont les mots synonymes de *m* dans WordNet auxquels aucune ressource n'est associée par Map_1 ($M^- = \bigcup_{w \in snonym(m) \wedge \forall s \in S, w \notin Map(s)} \{w\}$).

q_1= [+circuit "electrical circuit" "electric circuit" "electrical device" –tour -"racing circuit" -lap -circle]

q_2= [+circuit tour journey journeying -"electrical circuit" -"electric circuit -"electrical device" -"racing circuit" -lap -circle]

q_3= [+circuit path route itinerary -"electrical circuit" -"electric circuit" -"electrical device" -tour -"racing circuit" -lap -circle]

q_4= [+circuit group grouping -"electrical circuit" -"electric circuit" -"electrical device" -tour -"racing circuit" -lap -circle]

q_5= [+circuit "racing circuit" racetrack racecourse raceway track -"electrical circuit" -"electric circuit" -"electrical device" -tour -lap -circle]

q_6= [+circuit lap circle locomotion travel -"electrical circuit" -"electric circuit" -"electrical device" -tour -"racing circuit" -lap -circle]

Figure 4.13: Exemple de requêtes composées par le système de Santamarìa *et al.* pour le mot *m=circuit*

4.9 Bilan

Dans ce chapitre, nous avons présenté une approche d'extraction des exemples de résolution innovante de problèmes. Cette extraction est basée sur une désambigüisation sémantique des termes représentant les ressources d'innovation. Pour permettre cette désambigüisation nous utilisons une base de connaissances d'innovation qui permet de représenter l'association sémantique entre ces ressources et les différents opérateurs d'innovation.

Dans notre travail, nous nous sommes basés sur une ontologie d'innovation pour décrire les entités à localiser (i.e. les opérateurs et les ressources), les entités à extraire (i.e. les exemples) et leurs relations sémantiques.

L'objectif de cette description par ontologie est de faciliter l'adaptation de notre système d'extraction à d'autres domaines de connaissances ainsi qu'à d'autres méthodes d'innovation. Le système d'extraction se base sur cette ontologie pour reconnaitre automatiquement les

entités sémantiques alimentées dans la base de connaissances et extraire d'autres entités (i.e. les exemples) décrites dans cette ontologie.

Dans un domaine de connaissances, un expert se base sur cette ontologie pour l'enrichir par de nouveaux opérateurs et de nouvelles ressources et les mapper aux termes correspondants. Il peut aussi modifier l'ontologie (supprimer et ajouter des concepts et des relations) pour l'adapter à son domaine de connaissances. Cette modification peut nécessiter une mise à jour des règles d'extraction (qui assurent la désambiguïsation) pour adapter le système à la nouvelle ontologie.

Chapitre 5

Implémentation et expérimentations

Dans ce chapitre, nous présentons une mise en œuvre d'un prototype de notre approche. Nous proposons tout d'abord un module pour permettre à l'expert de définir dans une base de connaissances les opérateurs d'innovation existant dans son domaine de connaissances, tout en respectant l'ontologie d'innovation. Puis, la mise en correspondance (le mapping) des ressources des opérateurs à des termes d'une langue naturelle LN est faite par des tables relationnelles. Ensuite, nous présentons le fonctionnement du prototype implémenté. Nous terminons par une évaluation expérimentale de ce prototype et par une analyse critique du comportement observé.

5.1 Aide à l'élaboration de la base de connaissances

Pour séparer la base de connaissances de l'application et pour permettre à l'expert de représenter les opérateurs et les ressources dans un environnent indépendant [123], l'ontologie d'innovation O_{ino} (cf. § 4.2) a été traduite dans un schéma RDFS (*Resource Description Framework Schema* [122]). Ce schéma est présenté dans la figure 5.1. L'expert peut utiliser l'éditeur Protégé [124] pour définir dans un fichier RDF les opérateurs et les ressources spécifiques à son domaine de connaissances. La figure 5.2 présente une traduction en RDF (*Resource Description Framework* [125]) de l'opérateur loi de Coulomb (cf. § 4.3 et figure 4.6). Les codes RDFS et RDF présentés respectivement dans les figures 5.1 et 5.2 sont automatiquement produits à partir d'une description graphique dans l'éditeur Protégé (cf. figure 5.3). L'expert n'a donc pas à apprendre les langages RDFS et RDF pour pouvoir représenter les opérateurs et les alimenter dans la base de connaissances. Il lui faut seulement s'adapter à l'environnent de Protégé.

Figure 5.1: Codage de l'ontologie en RDFS

Protégé permet de représenter aussi l'ontologie avec d'autres langages et notamment OWL (*Web Ontology Language* [126]). Les différents niveaux de OWL (OWL Full, OWL DL, OWL Lite) sont plus expressifs que RDF(S). Mais ils sont plus difficiles à maîtriser et ne

sont pas adaptés à notre objectif (cf. § 4.1). L'expressivité de RDFS et RDF nous semble suffisante pour mettre au point respectivement notre ontologie et les bases de connaissances construites avec cette ontologie.

Dans la figure 5.1, chaque concept de notre ontologie (cf. § 4.2) est défini comme une classe RDFS. Cette définition commence par la balise <rdfs:Class> et finit par la balise </rdfs:Class>. D'autres balises de la forme <rdfs:subClassOf/> sont utilisées pour le rattacher à ses classes parents. Les attributs et les relations sont définis par les balises <rdf:property> et </rdf:property>. Les classes de départ et d'arrivée de chaque relation ou attribut sont encadrées par les balises <rdfs:domain/> et <rdfs:range/> respectivement.

Les attributs *about* et *label* des balises <rdf:Class> et <rdf:property> sont utilisés par Protégé pour donner respectivement une référence URI et un nom à chaque classe et relation définie. L'attribut *about* permet aussi de désigner les classes parentes et les propriétés par les balises <rdfs:subClassOf/>, <rdfs:domain/>, <rdfs:range/>.

En utilisant ce schéma (cf. figure 5.1), on peut définir une base de connaissances dans un fichier RDF (cf. figure 5.2). Dans cette figure, on représente l'opérateur loi de Coulomb défini en UML par le diagramme de la figure 4.6. Dans ce ficher, les balises <kb:[nom de classe]> </kb:[nom de classe]> sont utilisées pour définir les instances des classes du schéma RDFS. En plus des attributs *about* et *label* utilisés par Protégé, un troisième attribut *name* apparaît ici. C'est le nom que nous avons donné et utilisé pour désigner les ressources et les opérateurs de notre base de connaissances dans notre application. La balise <kb:[nom de relation] rdf:resource=[URI]/> est utilisée pour associer une instance à une autre repérée par l'attribut rdf:resource.

En utilisant l'éditeur de Protégé (cf. figure 5.3), l'expert peut facilement utiliser le schéma RDFS pour définir les opérateurs spécifiques à son domaine. Dans cette interface, il peut choisir une classe dans la fenêtre à gauche (i.e. *Effect*). Dans la fenêtre suivante, il définit une instance (i.e. *coulomb_law1*) et dans la dernière fenêtre il spécifie un nom par l'attribut *name* et le met en relation (i.e. *cause*, *improve*) avec d'autres instances dans la base de connaissances.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rdf:RDF (View Source for full doctype...)>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#
    xmlns:kb="http://protege.stanford.edu/kb#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  - <kb:Action rdf:about="http://protege.stanford.edu/kb#attraction"
      kb:name="attraction" rdfs:label="attraction">
      <kb:hypernym
        rdf:resource="http://protege.stanford.edu/kb#force" />
    </kb:Action>
    <kb:Action rdf:about="http://protege.stanford.edu/kb#charge"
      kb:name="charge" rdfs:label="charge" />
  - <kb:Effect
      rdf:about="http://protege.stanford.edu/kb#coulomb_law1"
      kb:name="coulomb_law" rdfs:label="coulomb_law1">
      <kb:cause
        rdf:resource="http://protege.stanford.edu/kb#charge" />
      <kb:improve
        rdf:resource="http://protege.stanford.edu/kb#repulsion" />
    </kb:Effect>
  - <kb:Effect
      rdf:about="http://protege.stanford.edu/kb#coulomb_law2"
      kb:name="coulomb_law" rdfs:label="coulomb_law2">
      <kb:improve
        rdf:resource="http://protege.stanford.edu/kb#attraction" />
      <kb:cause
        rdf:resource="http://protege.stanford.edu/kb#charge" />
    </kb:Effect>
  - <kb:Energy rdf:about="http://protege.stanford.edu/kb#force"
      kb:name="force" rdfs:label="force">
      <kb:hyponym
        rdf:resource="http://protege.stanford.edu/kb#attraction" />
      <kb:hyponym
        rdf:resource="http://protege.stanford.edu/kb#repulsion" />
    </kb:Energy>
    <kb:Substance
      rdf:about="http://protege.stanford.edu/kb#physical_object"
      kb:name="physical_object" rdfs:label="physical_object" />
  - <kb:Action rdf:about="http://protege.stanford.edu/kb#repulsion"
      kb:name="repulsion" rdfs:label="repulsion">
      <kb:hypernym
        rdf:resource="http://protege.stanford.edu/kb#force" />
    </kb:Action>
  </rdf:RDF>
```

Figure 5.2: Représentation sémantique de l'opérateur loi de Coulomb en RDF

Notre système d'extraction est implémenté en utilisant le langage de programmation JAVA. Les concepts, les attributs et les associations définis dans l'ontologie sont aussi décrits en JAVA. Ainsi, les instances définies dans un fichier RDF sont lues à l'amorçage du système et sont redéfinies comme instances des classes JAVA correspondantes. Dans la classe

Operator, nous avons défini trois procédures (*resource*, *object* et *example*) pour permettre la désambiguïsation, le raisonnement et l'extraction des exemples à partir des textes (cf. § 5.4.4).

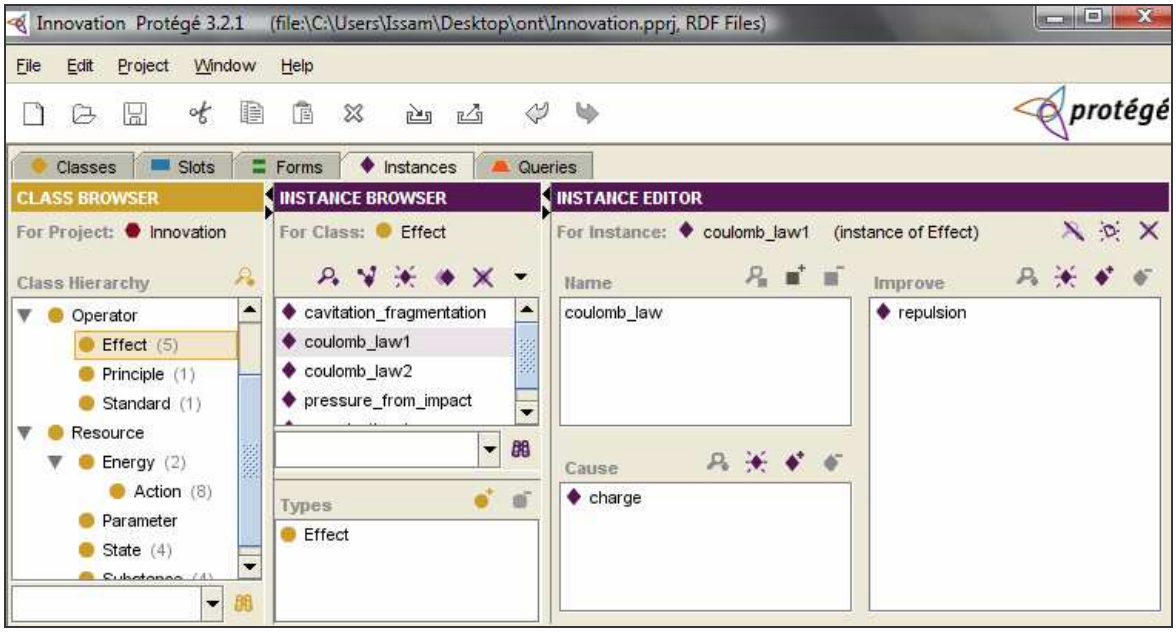


Figure 5.3: Interface graphique de l'éditeur de Protégé

5.2 Représentation des connaissances à extraire

En nous basant sur l'ontologie et les fonctions de mapping *Map*, Map_1 et Map_2 (cf. § 4.7.2), nous avons conçu un schéma UML (cf. figure 5.4) pour représenter les connaissances à extraire et à stocker dans la base de connaissances. Chaque fois qu'un exemple est ajouté, on le met en relation avec les ressources qui caractérisent l'opérateur.

Dans ce schéma, le concept *Example* de notre ontologie (cf. § 4.2.3) est associé à un autre concept *POS*. Ce dernier concept représente des informations syntaxiques de chaque mot dans le texte. Ses attributs *p_id*, *w_id*, *word*, *pos* représentent respectivement l'indice de la phrase, l'indice du mot, le mot du texte et sa catégorie lexicale (nom, verbe, adj, adv, préposition...). La forme canonique du mot (lemme) est définie par l'association avec le concept *Lemma* et le texte est défini par l'association avec le concept *Example*.

Un concept *Offset* représente un *synset* de WordNet. Son association avec le concept *Resource* définit le mapping Map_1 et son association avec le concept *Lemma* définit le

mapping Map_2. L'association entre les concepts *Ressource* et *Lemma* représente le mapping $Map = Map_1 \circ Map_2$.

Les concepts et les associations de ce schéma UML sont stockés dans une base de données relationnelle implémentée avec le système MySQL. La tâche de notre système est de remplir automatiquement les tables relationnelles une par une (cf. § 5.4.3). Seules les deux tables (*Map1Expert*, *MapExpert*) nécessaires pour implémenter le mapping Map_1 et *Map* sont remplies semi-automatiquement. Dans la section suivante, nous explicitons la partie semi-automatique du remplissage. Dans la section 5.4, nous nous basons sur les autres tables de la base pour présenter le fonctionnement du système et les étapes de leur remplissage.

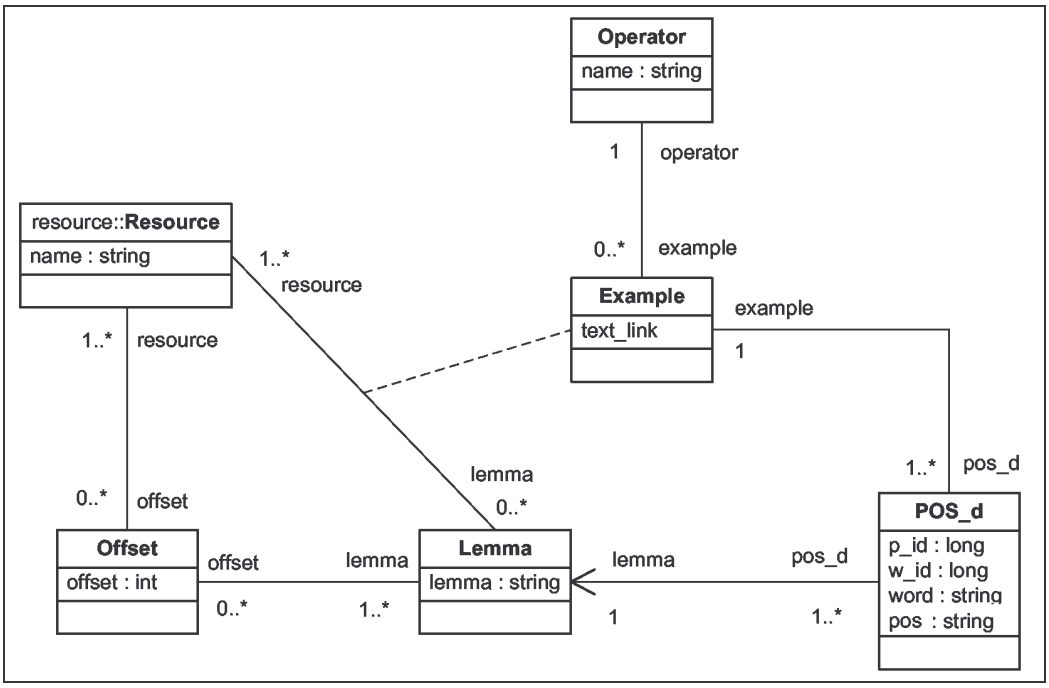


Figure 5.4: Schéma conceptuel des connaissances à extraire des textes en langue naturelle

Dans notre prototype, le recours à une base de données relationnelle est motivé par les raisons suivantes :

1. Faciliter les accès répétitifs aux connaissances et aux données qui sont nécessaires pour les traitements ultérieurs ;

2. Faciliter la représentation et la gestion de l'association lemme-ressource ;

3. Visualiser les connaissances extraites et recherche les défauts à corriger ;

4. Minimiser le temps nécessaire pour implémenter notre approche.

5.3 Mapping semi-automatique des ressources à des termes

Dans notre base de données, deux tables *Map1Expert* et *MapExpert* (cf. figure 5.5) sont définies pour permettre à l'expert de mapper les ressources de sa base de connaissances (un fichier RDF) respectivement à des *synsets* par la fonction Map_1 et à des termes par la fonction *Map* (cf. § 4.7).

Chaque ligne de la table *Map1Expert* associe une ressource à un *synset* de type *Noun* identifié dans WordNet par son *offset*. Dans cette table, l'expert peut associer une ressource à plusieurs *synsets* qui lui correspondent par Map_1 (i.e. $Map_1(charge) = \{8693651, 8809850, 10696423, ...\}$). Dès qu'un *synset n* est associé à une ressource *res*, implicitement tous les *synsets* hyponymes de *n* sont aussi associés par Map_1 à *res* et toutes les ressources hyperonymes de *res* sont aussi associées à *n*.

resouce	offset
physical_object	16236
substance	17572
pressure	105390
repulsion	903958
attraction	4493818
impact	6888434
shock_wave	6895233
charge	8693651
charge	8809850
attraction	10688069
repulsion	10688453
…	…

Table Map1Expert

resource	term
cavitation	cavitation

Table MapExpert

Figure 5.5: Association manuelle des ressources à leur synsets par Map_1 ou à leurs termes par Map

Le mapping Map_2 qui associe les *synsets* de la table *Map1Expert* à leurs termes et le mapping $Map = Map_1 \circ Map_2$ qui associe les ressources de cette table à leurs termes sont

définis implicitement dans notre système (cf. § 5.4.3). Ces mappings se basent sur WordNet, ses relations syntaxiques entre mots et ses relations sémantiques entre *synsets* (cf. § 4.7.1) pour mapper chaque lemme extrait du texte à ses offsets par Map_2 et ensuite à ces ressources par *Map.*

Dans le cas où un terme n'existe pas dans WordNet, la table *MapExpert* est prévue pour permettre à l'expert de l'associer explicitement à sa ressource par le mapping *Map* (i.e. le terme *cavitation* est associé directement à une ressource nommée aussi *cavitation* dans la base de connaissances). Cette table permet aussi d'intégrer d'autres langues que l'anglais dans le système i.e. le français).

5.4 Fonctionnement du système de recherche et d'extraction

Notre système de recherche et d'extraction fonctionne comme indiqué dans la figure 5.6.

Tout d'abord, le module de recherche documentaire sur le Web (cf. § 5.4.1) se base sur l'approche proposée dans le chapitre précédent pour repérer des documents candidats sur le Web.

Puis, si l'ensemble *D* de documents retournés par le moteur de recherche (i.e. Google.) n'est pas vide, on commence le traitement sur un document *d*. Sinon et s'il n'y a pas une demande d'arrêt de la part de l'expert, on sollicite le module de recherche pour effectuer une nouvelle recherche sur le Web.

Le traitement sur un document *d* commence par une analyse lexicale (cf. § 5.4.2). Ensuite, le système remplit les tables de la base relationnelle (cf. § 5.4.3). Le traitement s'achève par la désambiguïsation (cf. § 5.4.4), la suppression dans les tables des associations non pertinentes et le remplissage d'une table *d_exemple* représentant les associations à extraire.

Dès que le système achève le traitement du document *d*, il le supprime et il peut être arrêté par l'expert. Sinon il continue avec un prochain document de l'ensemble *D* de documents retournés par le système de recherche.

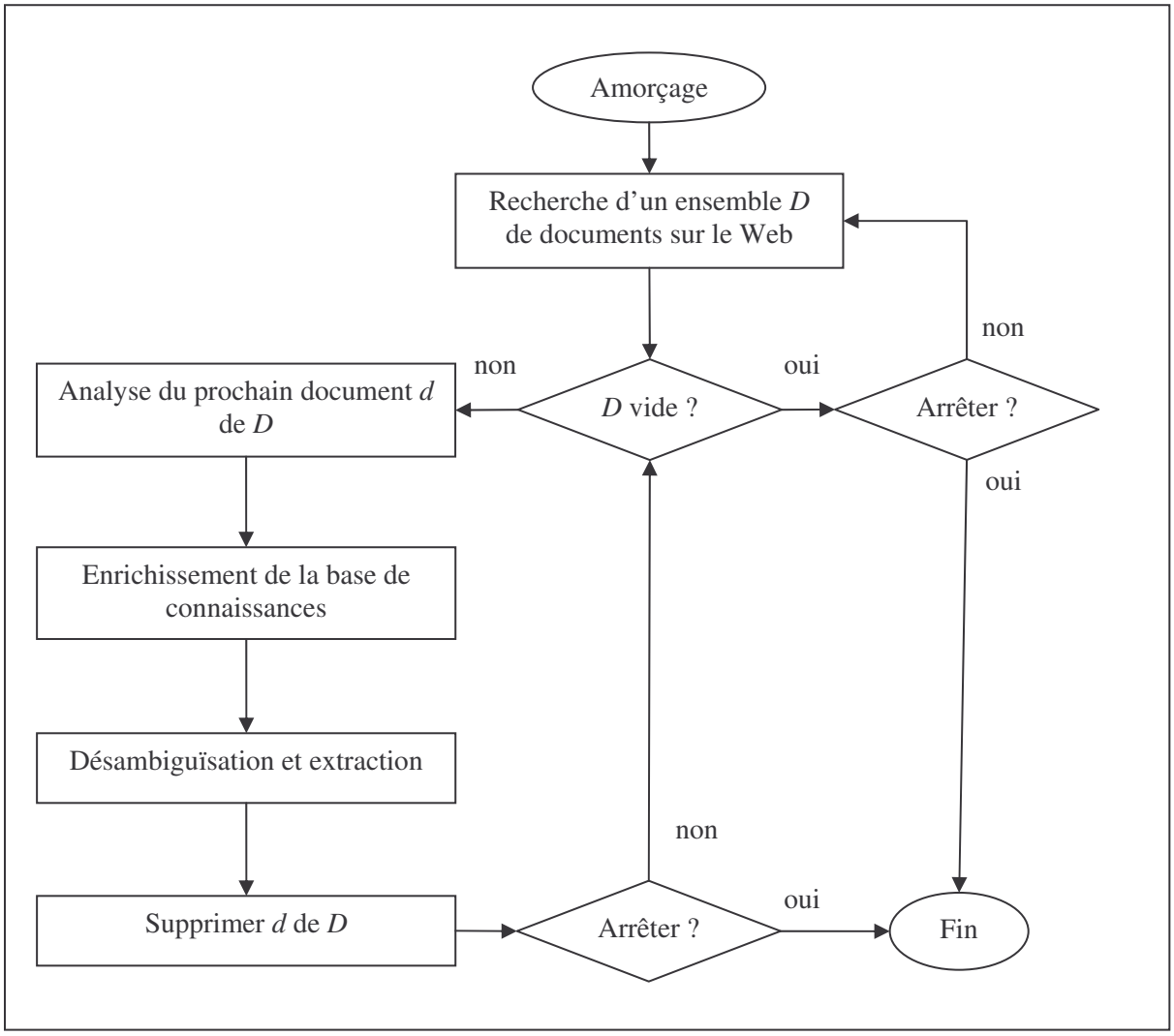


Figure 5.6: Fonctionnement du système de recherche et d'extraction

5.4.1 Recherche des documents sur le Web

Comme présenté dans le chapitre 3, c'est un module séparé. Son rôle est de composer des requêtes de recherche sur le Web par des mots clés mappés par *Map* à des ressources dans la base de connaissances. Une approche de sélection des mots est présentée dans le chapitre 4.

En se basant sur cette approche, pour chaque *synset n* de type *Noun* dans WordNet, s'il est mappé par Map_1 à une ressource dans la table *Map1Expert*, le système compose une requête de recherche $q(n)$ (cf. figure 5.7). Dans cette requête, les mots clés positifs (optionnels) sont les mots du *synset n* et les mots clés négatifs sont les mots des *synsets* synonymes de *n* et leurs hyponymes qui ne sont mappés à aucune ressource. Cette requête peut retourner de *0* à *m* documents à traiter.

q(0696423)=[charge "electric charge" -billing -presentment -countercharge -care -tutelage -guardianship -"due care" -"ordinary care" -"reasonable care" -"foster care" -"great care" providence -"slight care" -mission -commission -"fool's errand" -encumbrance -"assessment incumbrance"]

q(10696753)=["electrostatic charge"]

Figure 5.7: Des requêtes composées par la recherche sur le Web

5.4.2 Analyse lexicale d'un document

Dans notre système, nous avons utilisé l'analyseur TreeTagger [105, 127]. Cet analyseur est un projet développé à l'université de Stuttgart. Il existe des versions pour plusieurs langues dont le français et l'anglais. Il accepte en entrée une phrase ou un fichier texte et donne en sortie l'analyse lexicale. La figure 5.8 présente l'analyse effectuée pour la phrase « *During the machining process*, *a charge is applied to the abrasive wheel.* ».

Word	**POS**	**Lemma**
During	IN	during
The	DT	the
machining	VVG	machine
process	NN	process
,	,	,
a	DT	a
charge	NN	charge
is	VBZ	be
applied	VVN	apply
to	TO	to
the	DT	the
abrasive	JJ	abrasive
wheel	NN	wheel
.	SENT	.

Figure 5.8: Analyse syntaxique effectuée par l'analyseur TreeTagger

Cette analyse est représentée dans trois colonnes *word*, *POS* et *Lemma*. Dans notre système, cette analyse est effectuée sur chaque texte (document) et est alimentée dans un fichier temporaire *temp*. Ensuite, ce fichier est récupéré par le système et transféré dans la table *POS_d* (cf. § 5.4.3).

5.4.3 Remplissage des tables

Le remplissage commence par la table *POS_d* (cf. figure 5.9) dont le contenu représente le concept *POS* et ses associations au concept *Lemma* par la colonne *lemma* et au concept *Example* par la colonne *texte_link*.

text_link	p_id	w_id	word	pos	lemma
Abrasive_wheel.txt	1	1	During	IN	during
Abrasive_wheel.txt	1	2	the	DT	the
Abrasive_wheel.txt	1	3	machining	VVG	machine
Abrasive_wheel.txt	1	4	process	NN	process
Abrasive_wheel.txt	1	5	a	DT	a
Abrasive_wheel.txt	1	6	charge	NN	charge
Abrasive_wheel.txt	1	7	is	VBZ	be
Abrasive_wheel.txt	1	8	applied	VVN	apply
Abrasive_wheel.txt	1	9	to	TO	to
Abrasive_wheel.txt	1	10	the	DT	the
Abrasive_wheel.txt	1	11	abrasive	JJ	abrasive
Abrasive_wheel.txt	1	12	wheel	NN	wheel
Abrasive_wheel.txt	2	1	at	IN	at
Abrasive_wheel.txt	…	…	…	…	…

Figure 5.9: Table temporaire (*POS_d*) enrichi suite à l'analyse lexicale

Cette table est conçue pour mémoriser respectivement les données suivantes : le lien du document *d* (*text_link*), l'indice de la phrase (*p_id*) dans *d*, l'indice du mot (*w_id*), le mot (*word*), sa catégorie lexicale (*pos*) et le lemme (*lemma)*.

Deux autres tables *Map1_d* et *Map2_d* (cf. figure 5.10) sont utilisées pour associer les lemmes mémorisés dans la table *POS_d* à leurs ressources par la composition $Map = Map_1 \circ Map_2$. La table *Map1_d* représente l'association entre le concept *Offset* et le concept *Resource* et l'autre table *Map2_d* représente l'association des concepts *Offset* et *Resource*. Dans un premier temps, la table *Map2_d* est alimentée à partir du document *d* par le mapping Map_2. Ce mapping associe chaque lemme *lm* (nom ou verbe) dans la table *POS_d* à ses *synsets* de types *Noun* dans WordNet (cf. § 4.7.1). Si le lemme *lm* est un nom (la colonne *pos* dans la table *POS_d* doit contenir NN, NNS ou NP), l'association consiste à mettre en correspondance le lemme *lm* avec les offsets de ces *synsets* du type *Noun*. Si le lemme *lm* est un verbe (la colonne *pos* doit contenir VV, VVN, VBZ, VVG, VVD ou VVP), on l'associe

aux *synsets* de type *Noun* associés à ces *synsets* de type *Verb* par la relation *nominalisation.* La relation *nominalisation* est une abréviation de la relation qui associe les *synsets* de type *Noun* aux *synsets* du type *Verb* (et l'inverse) par l'intermédiaire de la relation lexicale *related_form* (cf. § 4.7.1).

resource	offset
repulsion	903958
physical_object	2596347
artefact	2596347
physical_object	2636079
artefact	2636079
physical_object	2734941
artefact	2734941
physical_object	2900890
artefact	2900890
artefact	3189674
physical_object	3189674
…	…

Table Map1_d

lemma	offset
repel	903958
surface	2596347
apply	2636079
wheel	2734941
charge	2900890
material	3189674
keep	3395681
machine	3561924
machine	3563571
grind	3585513
material	3590840
…	…

Table Map2_d

Figure 5.10: Tables nécessaires pour associer les lemmes d'un texte à leurs ressources par le mapping *Map*

La deuxième table *Map1_d* associe les *synsets* de la table *Map2_d* à leurs ressources, en se basant sur le mapping Map_1. Dans cette table, chaque *synset* associé à une ressource *res* est associé aussi à toutes les ressources hyperonymes de *res.* Et pour chaque *synset n* associé à une ressource *res*, les hyperonymes de *n* sont aussi associés à *res.*

```
INSERT INTO Map_d (resource, lemma)
   SELECT Map1_d.resource, Map2_d.lemma
   FROM Map1_d INNER JOIN Map2_d USING
(offset);
```

Figure 5.11: Requête retournant la table *Map_d* des ressources et des termes du document *d* associés par *Map*

La jointure (cf. figure 5.11) des tables *Map1_d*, *Map2_d* par leur colonne commune *offset* retourne une table *Map_d* contenant les ressources et les lemmes du document *d* associés par le mapping *Map*.

5.4.4 Désambiguïsation et extraction des exemples

La désambiguïsation et l'extraction s'achève par le filtrage à partir de *Map_d* (5.11) des associations qui correspondent à un exemple d'opérateur et leur transfert dans une table *t_example* (cf. figure 5.14).

Notre approche de désambiguïsation et d'extraction est implémentée comme indiqué ci-après.

La jointure (cf. figure 5.12) par la colonne commune *lemma* de la table *Map_d* et de la table *POS_d* (cf. figure 5.9) est utilisée pour remplir une table temporaire *POS_Resource* (cf. figure 5.13). Cette table est celle utilisée pour la désambiguïsation et pour l'extraction des exemples.

```
INSERT INTO POS_Resource (text_link, p_id, w_id, word, pos,
lemma, resource)
        SELECT POS_d.*, Map_d .resource
        FROM POS_d INNER JOIN Map_d USING (lemma);
```

Figure 5.12: Requête retournant une table temporaire *POS_Resource*

Pour cette tâche de désambiguïsation, les procédures (*resource*, *object*, *example*) de chaque opérateur *p* représenté par un objet JAVA sont utilisées.

Tout d'abord, la procédure *object* prend une copie *t_object* de la table *POS_Resource* et en supprime les associations syntaxiquement redondantes (cf. § 5.4.5). Puis, il supprime aussi toutes les lignes qui ne représentent pas des ressources apparues dans le contexte de l'opérateur *p*.

text_link	p_id	w_id	word	pos	lemma	resource
Abrasive_wheel.txt	1	6	charge	NN	charge	physical_object
Abrasive_wheel.txt	1	6	charge	NN	charge	charge
Abrasive_wheel.txt	1	6	charge	NN	charge	artefact
Abrasive_wheel.txt	1	12	wheel	NN	wheel	physical_object
Abrasive_wheel.txt	1	12	wheel	NN	wheel	artefact
Abrasive_wheel.txt	2	11	source	NN	source	physical_object
Abrasive_wheel.txt	2	11	source	NN	source	artefact
Abrasive_wheel.txt	2	16	workpiece	NN	workpiece	physical_object
Abrasive_wheel.txt	2	16	workpiece	NN	workpiece	artefact
Abrasive_wheel.txt	2	22	charged	VVN	charge	charge
Abrasive_wheel.txt	3	2	particles	NNS	particle	physical_object
Abrasive_wheel.txt	3	6	material	NN	material	physical_object
Abrasive_wheel.txt	…	…	…	…	…	…

Figure 5.13: Table *POS_Resource* utilisée pour l'extraction des exemples

Par exemple, pour un opérateur *p* d'un type effet (*Effect*), pour chaque ressource *res* associée par une relation *improve* ou une relation *cause* à *p* la requête relationnelle de la figure 5.15 est exécutée. Elle extrait de la table *POS_Resource* la liste de tous les indices des phrases (*p_id*) de la table *POS_Resource* qui correspondent à la ressource *res.* Elle supprime ensuite toutes les lignes de la table *t_object* dont le *p_id* n'est pas dans cette liste. La condition (*t_object.lemma<>POS_Resource.lemma*) est ajoutée pour empêcher la comparaison des lignes représentant un même lemme dans les deux tables *t_object* et *POS_Resource*.

text_link	lemma	resource	operator
Abrasive_wheel.txt	repel	repulsion	coulmb_law
Abrasive_wheel.txt	charge	charge	coulmb_law
Abrasive_wheel.txt	particle	substance	coulmb_law
Abrasive_wheel.txt	material	artefact	coulmb_law
Abrasive_wheel.txt	wheel	artefact	coulmb_law
Abrasive_wheel.txt	surface	artefact	coulmb_law

Figure 5.14: Table *t_example* associant les ressources à leurs lemmes par un opérateur d'innovation

```
DELETE t_object.* FROM t_object
 WHERE p_id NOT IN (
      SELECT DISTINCT POS_Resource.p_id
      FROM POS_Resource
      WHERE POS_Resource.type= res.name
      AND t_object.lemma <>
      POS_Resource.lemma
);
```

Figure 5.15: Requête pour supprimer à partir d'une table temporaire *t_object* les lignes qui ne représentent pas des ressources apparues dans le contexte d'une ressource *res* associée

Ensuite, la procédure *resource* de l'opérateur *p* à travers la requête de la figure 5.16 alimente la table *t_resource* correspondant à l'opérateur *p*. Cette requête est exécutée pour chaque ressources *res* associée à l'opérateur *p*. Elle sélectionne dans la table *POS_Resource* les lignes correspondant à la ressource *res* dont l'indice *p_id* (le contexte) correspond à des ressources dans la table *t_object* et elle les insère dans la table *t_resource.*

```
INSERT INTO t_resource (text_link, p_id, w_id, word, pos,
lemma, resource)
         SELECT * FROM POS_Resource
         WHERE POS_Resource.resource= res.name AND
         POS_Resource.p_id IN (
             SELECT DISTINCT t_object.p_id FROM t_object
             WHERE t_object.lemma <> POS_Resource.lemma
         );
```

Figure 5.16: Requête pour insérer dans une table temporaire *t_resource* les lignes de la table *POS_Resource* qui représentent une ressource *res* associée à un opérateur *p*

Puis, la procédure *example* insère le contenu des tables *t_object* et *t_resource* dans la table *t_example*. On obtient alors le résultant présenté dans la figure 5.17. Cette table *t_example* représente le concept *Example* qui associe les deux concepts *Resource* et *Lemma* et qui est associé au concept *POS_d* (cf. figure 5.4).

Après le remplissage de cette table *t_example a*vec des exemples pour tous les opérateurs de la base de connaissances, les associations sémantiquement redondantes sont supprimées par une procédure *delRedundancy* (cf. § 5.4.5). On obtient le résultat représenté dans la table de la figure 5.14.

text_link	lemma	resource	operator
Abrasive_wheel.txt	repel	repulsion	coulmb_law
Abrasive_wheel.txt	charge	charge	coulmb_law
Abrasive_wheel.txt	particle	physical_object	coulmb_law
Abrasive_wheel.txt	material	physical_object	coulmb_law
Abrasive_wheel.txt	particle	substance	coulmb_law
Abrasive_wheel.txt	material	artefact	coulmb_law
Abrasive_wheel.txt	wheel	physical_object	coulmb_law
Abrasive_wheel.txt	wheel	physical_object	coulmb_law
Abrasive_wheel.txt	surface	physical_object	coulmb_law
Abrasive_wheel.txt	surface	artefact	coulmb_law

Figure 5.17: Table *t_example* associant les lemmes du texte à leurs ressources candidates

5.4.5 Suppression des associations lemme-ressource redondantes

Deux types de redondances dans les associations lemme-ressource sont possibles dans la table *t_example*: la redondance sémantique et la redondance syntaxique.

1. La redondance sémantique peut exister lorsqu'un lemme est associé à des ressources hyperonymes. Cette redondance est supprimée de la table *t_example* par un post-traitement. Elle n'est pas supprimée plus tôt, car ces associations peuvent être nécessaires pour le repérage des opérateurs. La suppression est effectuée par une procédure *delRedundancy* définie comme suit :

   - Si un lemme *lm* est associé à une ressource *res* d'un opérateur alimenté dans la table *t_example*, toutes les autres associations du lemme *lm* avec les autres ressources sont considérées redondantes et sont supprimées. Par exemple, le lemme "*attraction*" est une force d'attraction selon la loi de coulomb. Dans un exemple de cet opérateur, l'association de ce lemme à la ressource *attraction* est acceptée et toute autre association est redondante et supprimée.

   - Sinon le lemme *lm* est associé à plusieurs ressources hyperonymes, les associations avec les ressources les plus spécifiques sont mémorisées et les autres sont supprimées. Par exemple dans la table *t_example* (cf. figure 5.14) toutes les associations à la ressource *physical_object* dans la figure 5.17 sont supprimées, si le lemme correspondant (i.e. "*wheel*", "*surface*", "*particle*") est

associé aussi à une ressource hyponyme de *physical_object* comme *artefact*, *substance*.

> An electrostatic dust precipitator comprising: a housing including an internally formed air path communicating between an inlet for drawing polluted air containing fine particulate and an outlet for emitting clean air purified by removing the fine particulate; …; a discharging means that is provided in the air path and that charges the fine particulate contained in the polluted air by using a corona discharge; a collecting means that is provided in the air path and that collects the fine particulate, which has been charged by the discharging means, by using a coulomb force; …. ; causes the fine particulate charged by the discharging means to repel from the charge surfaces; an attracting unit that includes an attraction shaft and a plurality of attracting plates,…
>
> Furuta *et al.* United States Patent N° 7297185, **http://www.uspto.gov/index.html**

Figure 5.18: Une partie d'un texte d'expérimentation contenant des mots représentant des ressources redondantes et utilisant le point-virgule à la place du point à la fin des phrases

2. La redondance syntaxique est représentée dans le text*e* de la figure 5.18 par des lemmes comm*e* *dust*, *cororna*, *attract*, *attraction et coulomb*, car ils sont utilisés comme modifieurs décrivant respectivement les mot*s* *precipitator*, *unit*, *discharge*, *force* qui représentent les ressources. Ces modifieurs sont des ressources pertinentes non redondantes s'ils apparaissent dans le texte sans les ressources qu'ils décrivent. Par exemple, le mo*t* *air* représente une ressource pertinente, car il apparait sans le mo*t* *path*. Ces redondances sont supprimées uniquement dans la table temporaire *t_object* par la requête de la procédure *object* (figure 5.19). On les maintient dans la table *POS_Resource*, car elles peuvent encore être nécessaires pour le repérage des opérateurs. Par exemple, dans le texte, si on supprime les lemmes *attract* et *attraction*, on risque de ne pas pouvoir extraire un exemple de l'opérateur loi de Coulomb.

```
DELETE t_object.* FROM t_object WHERE (w_id+1,
p_id) IN (
    SELECT DISTINCT w_id, p_id FROM POS_Resource
    WHERE t_object.p_id= POS_Resource.p_id
  );
```

Figure 5.19: Requête SQL utilisée pour supprimer la redondance syntaxique.

La requête de la figure 5.19 supprime toutes les lignes de la table *t_object* dont les indices *w_id+1* du mot suivant et *p_id* de la phrase représentent une ressource dans la table *POS_Resource*.

5.5 Expérimentations

5.5.1 Elaboration d'une base de connaissances:

Nous avons spécifié une base de connaissances (cf. annexe G) pour quatre opérateurs d'innovation et onze ressources. Ces opérateurs du type effet sont les suivants: loi de Coulomb, fragmentation de liquides par cavitation, pression de l'impact liquide-solide, évaporation de liquides en diminuant la pression de sa vapeur saturée. Ces opérateurs sont présentés avec des exemples dans l'annexe E. Ils sont décrits par les ressources *charge*, *attraction*, *repulsion*, *force*, *cavity*, *pressure*, *vaporization*, *wave*, *impact.*

5.5.2 Préparation de la collection

Text_link	**Termes annotés**	**Opérateur correspondant**
static.html	electron, charge, stick, wool, balloon, hair, door	coulomb_law
Goldfire Innovator	rock, material, water, pressure, shoch_wave	shock_wave
5473787.html	tube, exchanger, lance, wave cleaning, water, pressure, shock	shock_wave
vapor_pressure	water, vapor, pressure	vaporization _by_pressure
5947784.html	bubble, fluid, air, valve, surface	cavitation_ fragmentation
atmospheric_ electricity	-	-

Figure 5.20: Tableau utilisé pour l'annotation manuelle de la collection

Les textes sont recherchés dans diverses sources : manuellement sur le Web, dans la base de connaissances de Goldfire Innovator (cf. § 1.3.1), sur le site américain de brevets USPTO[9]

[9] http://www.uspto.gov/

(*United States Patent And Trademark Office*), sur le site Patent Storm[10]. Nous avons préparé une collection (cf. annexe H) de soixante-trois (63) textes. Dans cette collection, nous avons manuellement annoté cent quatre-vingt-sept (187) termes représentant des ressources objets de soixante-dix exemples pertinents. Cette taille modeste de la collection s'explique par la difficulté du repérage et de l'annotation des textes.

Pour évaluer notre approche, nous avons repéré les exemples dans la collection. Le lien vers l'exemple et l'opérateur correspondant sont annotés dans la figure 5.20. Nous avons repéré dans chaque exemple les lemmes qui représentent des ressources pertinentes (cf. figure 5.20). Suite à cette annotation, les nouvelles ressources apparues (*physical_phenomenon*, *physical_object*, *artefact*, *substance*, *solid*, *plasma*) sont ajoutées dans la base de connaissances (cf. annexe G) et sont ensuite mappées à leurs termes par les tables *Map1Expert*, et *MapExpert* (cf. figure 5.5).

5.5.3 Évaluation

Sur cette collection ainsi préparée, on a expérimenté le système qui nous a fourni les exemples et les ressources repérées. Certain de ces exemples et ressources sont pertinents et d'autre ne sont pas pertinents. Nous avons ainsi pu calculer la précision *P*, le rappel *R* et leur combinaison *P&R* (cf. chapitre 2). Les scores obtenus sont présentés dans la figure 5.21.

	P	**R**	**P&R**
Exemple	81,33%	87,14%	84,14%
ressources	73,36%	88,18%	80,09%

Figure 5.21: Scores réalisés dans nos expérimentations

Ce tableau montre que notre approche est plus performante pour l'annotation des exemples que pour l'annotation de leurs ressources. Cette différence entre l'annotation des ressources et des exemples s'explique par plusieurs raisons :

1. Dans notre collection, des exemples pertinents (cf. figure 5.22) ne sont pas repérés.

[10] http://www.patentstorm.us/

If you charge a balloon by rubbing it on your hair, it picks up extra electrons and has a negative *charge*. Holding it near a neutral object will make the charges in that object move. If it is a conductor, many electrons move easily to the other side, as far from the balloon as possible. If it is an insulator, the electrons in the atoms and molecules can only move very slightly to one side, away from the balloon. In either case, there are more positive charges closer to the negative balloon. Opposites attract. The balloon sticks. (At least until the electrons on the balloon slowly leak off.) It works the same way for neutral and positively charged objects.

http://www.sciencemadesimple.com/static.html

Figure 5.22: Un exemple pertinent non repéré dans un texte d'expérimentation

Pour ce cas le système n'a pas trouvé de nouvelles ressources dans le contexte de l'opérateur (dans la figure 5.22, le mot *stick* n'est pas compris comme *attraction*, faute de mapping). Des exemples ainsi que des ressources impertinents peuvent être extraits par le système. Ces cas sont surtout constatés pour des documents qui présentent un opérateur d'innovation par des ressources très générales (cf. figure 5.23) comme (*thing*, *object*). Si ces mots généraux sont mappés à des ressources comme *physiqual_object* par *Map*, le système les prend en compte.

Now, positive and negative charges behave in interesting ways. Did you ever hear the saying that opposites attract? Well, it's true. Two **<res>things</res>** with opposite, or different **<res>charges</res>** (a positive and a negative) will **<res>attract</res>,** or **<res>pull</res>** towards each other. **<res>Things</res>** with the same **<res>charge</res>** (two positives or two negatives) will <res>repel</res>, or push away from each other.

http://www.sciencemadesimple.com/static.html

Figure 5.23: Un exemple non pertinent repéré dans un texte d'expérimentation

2. Le système se base sur les opérateurs et leur contexte pour annoter les ressources et ne prend pas en compte les informations syntaxiques et sémantiques pour repérer les ressources objets. Dans la figure 5.24, on présente une partie d'un document dans la collection d'expérimentation qui est un exemple de l'opérateur loi de Coulomb. Les termes annotés automatiquement par le système sont encapsulés dans les deux balises <res> et </res> et les termes manuellement annotés sont soulignés. Dans ce texte, on remarque la redondance des mots annotés par le système. Elle provient du manque

d'information syntaxique et sémantique sur les ressources objets des opérateurs. Ces ressources qui sont extraites du contexte d'un opérateur ne sont pas définies dans la base de connaissances par des relations sémantiques avec cet opérateur. On ne connaît pas à l'avance leur type. Elles ne sont définies par aucune information syntaxique. Par conséquent des mots impertinents dans le mapping (i.e. *type*, *work* de la figure 5.24) pour représenter un exemple peuvent leur être associés, car ils apparaissent dans le contexte des ressources de l'opérateur.

Because the two **<res>types</res>** of ice **<res>fragments</res>** have opposite **<res>charges</res>**, they **<res>attract</res>** each other: but **<res>gravity</res>** **<res>pulls</res>** the bigger ones down, while the **<res>wind</res>** blows smaller slivers even higher, and in separating the two **<res>types</res>**, these two **<res>forces</res>** perform <res>work</res> against the electric **<res>attraction</res>.** The situation is therefore somewhat similar to Robert Van de Graaff's machine, except that there the rubber band overcomes electric **<res>repulsion</res>**, while here, the **<res>forces</res>** of the **<res>wind</res>** and of **<res>gravity</res>** overcome an **<res>attraction</res>**. Still, **<res>work<res>** is **<res>work</res>**, and by performing it the process increases the energy stored in the system. The top of the cloud where the little slivers end up, becomes **<res>charged</res>** to a high positive_voltage, until the air cannot contain the growing electric **<res>charge</res>** any more, and... FLASH! BOOOOM!

http://www.iki.rssi.ru/mirrors/stern/stargaze/Svandgrf.htm

Figure 5.24: Partie d'un document pertinent de la collection d'expérimentation (les annotations manuelles sont soulignées et les annotations automatiques sont encapsulées dans des balises)

3. L'existence de fautes et de symboles dans les textes qui échappent à l'analyse lexicale. Ces fautes induisent une analyse et une annotation erronées dans les textes. Un symbole dans un texte peut être annoté comme nom. Des lettres (i.e. *h* ou *g*) peuvent représenter des mots dans WordNet. Par conséquent ils représentent des ressources candidates dans le texte. Dans un document d'expérimentation[11], la phrase « *So*, *you've seen a field before*, *in the form of g. Electric fields operate in a similar way.* » est mal analysée, à cause de l'existence de la lettre *g*. L'analyseur a considéré que les deux phrases du texte constituent une seule phrase. La lettre *g* avec le point « *g.* » a été reconnu comme un adjectif (*pos* est JJ dans la table *POS_d* cf. figure

[11] http://physics.bu.edu/~duffy/PY106/Charge.html

5.25) ; le mot *Electric* a été reconnu comme un nom propre (*pos* est NP). Le mot *Electric*, à partir du *synset* nominal {*electric*, *electric automobile*, *electric car* -- (*a car that is powered by electricity*)} est alors associé à la ressource *artefact* de notre base de connaissances. Si on supprime le point, l'analyseur reconnaît *g* comme nom et si on remplace *g* par un mot anglais, l'analyseur fait une analyse correcte.

text_link	p_id	w_id	word	pos	lemma
http://.../harge.html	…	…	…	…	…
http://.../harge.html	84	7	in	IN	in
http://.../harge.html	84	8	the	DT	the
http://.../harge.html	84	9	form	NN	form
http://.../harge.html	84	10	of	IN	of
http://.../harge.html	84	11	g.	JJ	g.
http://.../harge.html	84	12	electric	NP	electric
http://.../harge.html	84	13	fields	NNS	field
http://.../harge.html	…	…	…	…	…

Figure 5.25: Analyse erronée résultant de l'existence d'abréviations dans le texte

Sur un autre texte[12], le test (cf. figure 5.26) nous a donné comme ressources les lettres (*e*, *g* et *h*) qui sont utilisées pour des énumérations dans le texte.

text_link	lemma	resource	operator
5817374.html	adhere	attraction	coulomb_law
5817374.html	attraction	attraction	coulomb_law
5817374.html	pull	attraction	coulomb_law
5817374.html	charge	charge	coulomb_law
5817374.html	bed	artefact	coulomb_law
5817374.html	particle	physical_object	coulomb_law
5817374.html	mask	artefact	coulomb_law
5817374.html	e	substance	coulomb_law
5817374.html	g	substance	coulomb_law
5817374.html	h	substance	coulomb_law
5817374.html	H.	substance	coulomb_law

Figure 5.26: Annotation impertinente des lettres par des ressources

[12] http://www.freepatentsonline.com/5817374.html

4. La mauvaise détermination des fins de phrases par l'analyseur lexical TreeTagger. Cet analyseur utilise des indicateurs comme ('.', '?', '!') pour déterminer la fin d'une phrase. Par conséquent, l'existence de ces caractères dans une autre position peut couper la phrase en deux. L'utilisation des autres marques (i.e. ';') ou non (i.e. les titres), provoque la jointure de la phase avec sa suivante.

Un texte de brevet (cf. figure 5.18) utilisé dans nos tests a beaucoup employé le point-virgule à la place du point. Sur deux paragraphes, le système a annoté 24 lemmes (cf. table I, figure 5.27) qui représentent des ressources. En remplaçant à nouveau le point-virgule par le point, les deux mots (*journal*, *attracting-plate*) ne sont plus sélectionnés (cf. table II figure 5.27)[13]; le premier n'est pas pertinent et le deuxième est redondant.

lemma	lemma
air	means
airstream	outlet
attract	path
attracting-plate	plate
attraction	precipitator
attraction-shaft	repel
charge	shaft
discharge	shape
draw	state
force	surface
housing	unit
journal	voltage

Table I: Lemmes extraits du texte avec point-virgule ';'

lemma	lemma
air	outlet
airstream	path
attract	plate
attraction	precipitator
attraction-shaft	repel
charge	shaft
discharge	shape
draw	state
force	surface
housing	unit
means	voltage

Table II: Lemmes extraits du texte avec point '.'

Figure 5.27: Réduction des lemmes annotés comme ressources par le remplacement d'un point-virgule par un point pour marquer les fins de phrases (dans les tables les mots impertinents sont soulignés)

Les mots *attracting-plate*, *attraction-shaft* sont redondants dans les tables de la figure 5.27, car les mots *plate*, *attraction* et *shaft* figurent déjà dans la tables. Ces mots ne sont pas supprimés, car l'analyseur lexical Treetagger n'a pas réussi à identifier leur sous lemmes.

[13] Le test sur le document complet d'environ 9500 mots nous a permis de réduire ce nombre de 92 à 83 ressources.

Dans notre évaluation les mots redondants (syntaxiquement ou sémantiquement) sont considérés comme impertinents s'ils ne représentent pas des ressources pertinentes.

Note : un mot (cf. *field*, *earth* figure 5.28) peut être associé à plusieurs ressources différentes qui ne sont pas liées par une relation d'hyperonymie (*artefact*, *substance*, *physical_phenomenon*). Nous pensons que notre approche va permettre de filtrer la (les) association(s) la (les) plus pertinente(s) lorsque la base de connaissances associe ces ressources à un nombre plus important d'opérateurs. Ainsi, dans notre évaluation ces mots sont considérés comme pertinents si une association pertinente figure dans la table.

text_link	lemma	resource	operator
charge17.txt	air	physical_phenomenon	coulomb_law
charge17.txt	air	substance	coulomb_law
charge17.txt	change	physical_object	coulomb_law
charge17.txt	conductivity	physical_phenomenon	coulomb_law
charge17.txt	construction	artefact	coulomb_law
charge17.txt	charge	charge	coulomb_law
charge17.txt	collection	artefact	coulomb_law
charge17.txt	collector	artefact	coulomb_law
charge17.txt	earth	artefact	coulomb_law
charge17.txt	earth	substance	coulomb_law
charge17.txt	field	artefact	coulomb_law
charge17.txt	feld	physical_phenomenon	coulomb_law
charge17.txt	…	…	…

Figure 5.28: Association des mots aux ressources candidates

5.6 Bilan

Dans ce chapitre nous avons présenté un prototype permettant l'extraction des exemples de résolution innovante de problèmes par notre approche présentée dans le chapitre 4. Ce prototype se base sur une représentation en RDF/ RDFS pour la mise en œuvre de la base de connaissances et de son ontologie. Les exemples à extraire sont stockés dans une base de données relationnelle. Ainsi les règles de désambigüisation et d'extraction sont formulées par des requêtes SQL encapsulées dans un code JAVA. Nous avons expérimenté ce prototype sur une collection de 63 textes de brevets. L'évaluation sur cette collection montre l'efficacité de

cette approche : (P&R=84,14%) pour l'extraction des exemples contre (80.09%) pour l'annotation de leurs ressources.

Pour ce prototype l'expert peut modifier la base de connaissances mais pas l'ontologie. La modification de l'ontologie nécessite en effet un langage pour permettre la mise à jour des règles d'extraction.

Conclusion et perspectives

Dans cette thèse, nous nous somme intéressés à la génération d'une base de connaissances pour le domaine de l'innovation et l'enrichissement permanent de cette base par des exemples de résolution inventive des problèmes d'innovation extraits à partir de textes en langue naturelle. L'objectif est de permettre l'automatisation de la résolution des problèmes d'innovation.

Dans un premier temps nous avons identifié les concepts nécessaires pour constituer cette base. Ces concepts sont représentés dans une ontologie d'innovation. Grâce à cette ontologie, un expert d'un domaine de connaissances peut définir dans la base des solutions générales d'innovation appelées opérateurs. Dans l'ontologie, trois types d'opérateurs sont définis : *Effect*, *Principle*, *Standard*. Ces types se différencient par les rôles qui les associent à des entités sémantiques appelées ressources d'innovation. Le concept *Effect* définit une relation cause-conséquence entre ressources. Le concept *Principle* définit une relation de contradiction entre des paramètres d'innovation et le concept *Standard* représente les solutions innovantes génériques qui résolvent les problèmes en améliorant les états de leurs composants. Les ressources sont représentées par un concept *Resource* et elles sont associées par des relations sémantiques comme la relation d'hyperonymie/hyponymie et la relation de partie/tout.

L'intérêt de cette base de connaissances est de pouvoir associer des exemples à chacun des opérateurs. Le principal problème que nous avons étudié est l'extraction automatique d'exemples à partir de textes issus du Web. Dans un premier temps nous avons étudié l'adaptation à ce problème de deux approches existantes : l'extraction d'informations à partir d'un formulaire, l'extraction d'informations à partir d'une requête utilisateur. Ces approches sont basées sur des techniques de traitements des langues naturelles et peuvent faire appel à l'apprentissage. Il nous est apparu qu'aucune des deux ne pouvait satisfaire complètement notre objectif. Elles se focalisent sur des entités nommées et dans notre application on doit se focaliser sur des ressources d'innovation définies dans la base de connaissances. On doit repérer leurs spécifications dans les textes et les utiliser pour extraire les exemples pour les opérateurs. L'adaptation d'une approche de TALN pour notre application nécessite une mise

au point manuelle longue et difficile pour élaborer les règles d'extraction. Les techniques d'apprentissage butent sur la constitution d'un corpus d'entrainement représentatif.

Par conséquent, la conception d'une nouvelle approche d'extraction adaptée à notre application s'est imposée. Cette approche comprend deux grandes tâches. La première tâche est le repérage des textes sur le Web. La deuxième tâche est l'extraction d'exemples pour les opérateurs à partir de ces textes. La première tâche est effectuée avec un moteur de recherche classique. Les requêtes de recherche sont composées de mots clés positifs et de mots clés négatifs. Les mots positifs sont les mots des *synsets* de WordNet mappés aux ressources de la base. Les mots négatifs sont les mots synonymes qui ne sont mappés à aucune ressource. Pour l'extraction des exemples à partir des textes, nous avons mis au point une technique spécifique réalisant la désambiguïsation du sens des mots. Un mapping des ressources de la base de connaissances à WordNet aide à la mise en œuvre de ces deux tâches. Dans la première tâche, il permet d'identifier les mots clés positifs et les mots clés négatifs. Dans la deuxième tâche, il permet de repérer dans les textes les mots qui représentent les ressources des opérateurs et qui nécessite une désambiguïsation.

Un prototype de notre approche a été implémenté en JAVA. Une base de données relationnelle est conçue pour contenir les exemples. Le filtrage des informations pertinentes pour chaque exemple est réalisé par l'intermédiaire de requêtes SQL. Une table dans cette base relationnelle est utilisée pour permettre le mapping des ressources de la base de connaissances à des *synsets* dans WordNet. Par l'intermédiaire de l'environnement de développement d'ontologie Protégé, l'expert peut définir sa base de connaissances en RDF.

Nous avons mené des expérimentations sur une collection de 63 textes, principalement issus du Web. Ces textes ont été manuellement annotés par leurs ressources et leurs opérateurs d'innovation. Une base de connaissances a été constituée avec quatre opérateurs d'innovation et des ressources d'innovation repérées dans la collection. Les évaluations menées sur la collection ont montré que notre approche est prometteuse, en réalisant le score P§R =84.14 % pour l'extraction d'exemples pour les opérateurs et le score P§R =80.09 % pour l'extraction des spécifications de ressources.

Néanmoins, ces expérimentations ne sont pas suffisantes pour déterminer complètement le comportement de notre approche. Il faudrait poursuivre les tests sur une collection plus importante de textes manuellement annotés. Il faudrait aussi définir une base de connaissances

plus riche avec un plus grand nombre d'opérateurs et de ressources d'innovation. Un problème que nous avons rencontré est celui de la pertinence de l'association des termes aux ressources. Il est résolu par notre technique de désambiguïsation et cette technique ne marche bien que si la base contient suffisamment de ressources, d'opérateurs et de relations entre eux. Ainsi, il serait intéressant d'évaluer l'effet de la taille de notre base et de la richesse des associations ressource-opérateur sur le comportement de notre approche.

Dans notre travail, pour des raisons de simplicité, les ressources de type fonction ne sont pas prises en compte. Une fonction met des ressources en relation sur la base de triplets (sujet, action, objet). Dans notre ontologie, les ressources sont directement associées à l'opérateur sans la fonction. Il serait intéressant d'étudier une structuration de la base de connaissances permettant de gérer les fonctions.

Le prototype implémenté et présenté dans le chapitre 5 peut être facilement adapté à d'autres ontologies, à d'autres domaines de connaissances, à d'autres méthodes d'innovation. Pour atteindre pleinement cet objectif, il lui manque seulement un langage permettant la mise à jour des règles d'extraction pour prendre en compte les modifications effectuées dans l'ontologie.

Dans cette thèse nous n'avons pas abordé la résolution automatique/semi-automatique des problèmes d'innovation à partir de notre base de connaissances. Pour vérifier la pertinence de notre proposition nous avons développé une interface utilisateur qui permet d'interroger la base de connaissances et d'accéder à un opérateur et à ses exemples par des hyper liens. Chaque exemple est affiché avec une annotation dynamique de ses ressources dans les textes. Il serait intéressant de concevoir un système de résolution automatique/semi-automatique de problèmes. Un langage de requête approprié pourrait permettre à l'utilisateur d'exprimer son problème. Dans cette optique, le système devrait être capable d'analyser les requêtes exprimées dans ce langage et d'identifier le problème et son opérateur dans la base de connaissances. En exploitant les exemples associés à cet opérateur dans la base de connaissances, il proposerait des solutions adaptées au problème.

Bibliographie

[1] G. Altschuller: *Creativity as an Exact Science: The Theory of the Solution of Inventive Problems*. Gordon and Breach Science Publishers, New York, 1984

[2] G. Altshuller: *TRIZ The innovation algorithm: systematic innovation and technical creativity*. Traduit par Lev Shulyak et Steven Rodman, Technical Innovation Centre, Worcester, 1999

[3] G. Gogu: *Méthodologie d'innovation: la résolution des problèmes créatifs*. Revue Française de Gestion Industrielle, Vol. 19, pp. 35-62, 2000

[4] G. Mazur: *Theory of Inventive Problem Solving (TRIZ)*. 1996
http://www.mazur.net/triz/

[5] J. Terninko, Z. Alla, Z. Boris: *Step-by-Step TRIZ: Creating Innovative Solution Concepts*. Responsible Management, Nottingham, 1996

[6] D. Choulier, G. Draghici: *TRIZ : une approche de résolution des problèmes d'innovation dans la conception de produits*. Modélisation de la connaissance pour la conception et la fabrication intégrées. Editura Mirton, pp. 31-58, 2000

[7] J. Terninko, A. Zusman, B. Zlotin: *Systematic Innovation: An Introduction to TRIZ*. CRC Press, 1998

[8] D. Cavallucci, P. Lutz: *TRIZ: un concept nouveau de résolution des problèmes d'innovation*. 2ème Congrès International Franco-Québécois de Génie Industriel, Albi, 1997

[9] C. Esteyries: *Gagner des avantages stratégiques par l'innovation systématique: La méthodologie I-TRIZ*. 2005
http://www.champagne-ardenne-tech.fr/-spip/article.php3?id_article=324

[10] Ideation International: *TRIZ & ITRIZ*. http://www.ideationtriz.com/triz.asp

[11] H. Linde, G. Herr, A. Rehklau: *Innovation of the Integrated Product and Process Development by WOIS*. TRIZ Conference, Osaka, 2006

[12] H. Linde, B. Hill: *Erfolgreich Erfinden: Widerspruchsorientierte Innovationsstrategie*. Darmstadt, Hoppenstedt, 1993

[13] E.N. Sickafus: *Unified Structured Inventive Thinking: How to Invent*. Ntelleck, Grosse Ile, 1997

[14] Y. Akao: *Development History of Quality Function Deployment*. The Customer Driven Approach to Quality Planning and Deployment, pp. 339-351, Tokyo, 1994

[15] QFD Institute: *QFD: Quality Function Deployment*. http://www.qfdi.org/

[16] A.F. Osborn: *Applied Imagination: Principles and Procedures of Creative Problem Solving*. Charles Scribner's Sons, New York, 1963

[17] K. Hunt: *CPS Model*. Notes from Gary Davis's Creativity is Forever, 1998
http://members.optusnet.com.au/~charles57/Creative/Brain/cps.htm

[18] Invention Machine corporation: *Goldfire Innovator.* http://invention-machine.com/GoldfireInnovator.htm

[19] Innovation WorkBench® 3.2. http://www.ideationtriz.com/new/iwb.asp

[20] TRIZ Explorer™. http://www.insytec.com/trizexplorer.htm

[21] V. Fey: *Glossery of TRIZ.* triz journal, 2001 http://www.triz-journal.com/archives/2001/03/a/

[22] Invention Machine corporation : *TechOptimiser3.5 Tutor*. 1995-2000

[23] *Matrice TRIZ interactive & 40 Principes*. http://www.triz40.com/aff_Matrice.htm

[24] P. Frescal, E. Portales, G. Gogu: *développement d'une interface d'aide pour TechOptimizer*. Institut Français de Méchanique Avancée (IFMA), 2005

[25] Applied Innvations Group: *R&D Techoptimiser software-overview.* http://www.applied-innovation.com/Products/RDInnovation/RDTechOptimiser/tabid/92/Default.aspx

[26] Knowllence : *CreaTRIZ*: *Méthode TRIZ et recherche de solutions*. http://www.knowllence.com/fr/produits/creatriz.php

[27] TRISolver 2.1™. http://www.triz-journal.com/archives/2002/05/h/index.htm

[28] F. Ibekwe-SanJuan : *Fouille de textes : méthodes, outils et applications*. Hermès, Paris, 2007

[29] T. Poibeau : *extraction automatique d'information : de texte brut au Web sémantique*. Hermès, Paris, 2003

[30] E. Riloff: *Automatically Generating Extraction Patterns from Untagged Text*. 13th National Conference on Artificial Intelligence (AAAI '96), pp. 1044-1049, Portland, 1996

[31] S. Soderland, D. Fisher, J. Aseltine, W. Lehnert: *Crystal: Inducing a Conceptual Dictionary*. 14th International Joint Conference on Artificial Intelligence (IJCAI-95), pp.1314-1319, Montréal, 1995

[32] F. Ciravegna: *Adaptive Information Extraction from Text by Rule Induction and Generalisation*. 17th International Joint Conference on Artificial Intelligence (JCAI 2001), pp. 1251-1256, Seattle, 2001

[33] D. Bikel, R. Sschwartz, R. Weischedel: *An algorithm that learns what's in a name*. Machine Learning, Vol. 34, pp. 211-231, Kluwer Academic Publishers, Hingham, 1999

[34] B. Yildiz, S. Miksch: *Motivating Ontology-Driven Information Extraction.* International Conference on Semantic Web and Digital Libraries (ICSD-2007), pp. 45–53, Bangalore, 2007

[35] E. SanJuan, J. Dowdall, F. Ibekwe-SanJuan, F. Rinaldi: *A symbolic approach to automatic multiword term structuring.* Computer Speech and Language (CSL), Vol. 19, N. 4, pp. 524-542, Elsevier, Oxford, 2005

[36] G. Krupka: *Description of the SRA system as used for muc-6*. 6th Message Inderstanding Conference (MUC-6), pp. 221-235, Maryland, 1995

[37] F. Ibekwe-SanJuan, E. anJuan: *TermWatch: cartographie de réseaux de termes*. 5th conference on terminology and Artificial Intellegence (TIA'03), Strasbourg, pp. 124-134, 2003

[38] B. Daille: *Conceptual structuring through term variations*. 41st annual meeting of the Association for Computational Linguistics (ACL 2003), pp. 9-16, Sapporo, 2003

[39] D. Faure, C. Nedellec: *Asium: learning subcategorization frames and restrictions of selection*. 10th European Conference on Machine Learning (ECML 98), Chemnitz, 1998

[40] B. Grau: *Méthodes avancées pour les systèmes de recherche d'informations*. Systèmes de question-réponse, pp. 189-218, Hermès, 2004

[41] W. Lehrent: *Human and computational question answering*. Cognitive Science, Vol. 1, pp. 47-63, 1977

[42] S. Harabagiu, M. Pasca, S. Maiorano: *Experiments with Open-4Domain Textual Question Answering*. 18th conference on Computational linguistics (COLING-2000), pp. 292-298, Saarbruken, 2000

[43] C. Jacquemin: *Spotting and discovering terms through NLP*. MIT Press, Cambridge, 2001

[44] M. Soubbotin, S. Soubbotin: *Use of patterns for detection of likely answer strings: A systematic approach*. 11th Text REtrieval Conference (TREC 2002), Gaithersburg, 2002

[45] D. Lin, P. Pantel: *Discovery of inference rules for question-answering*. Natural Language Engineering, Vol. 7, pp. 343-360, Cambridge University Press, New York, 2001

[46] D. Moldovan, D. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Badulescu, O. Bolohan: *LCC Tools for Question Answering*. 11th Text REtrieval Conference (TREC 2002), Gaithersburg, 2002

[47] S. Harabagiu, D. Moldovan, M. Bowden, C. Clark, A. Hickl, P. Wang: *Employing Two Question Answering Systems in TREC-2005*. 14th Text Retrieval Conference (TREC 2005), Gaithersburg, 2005

[48] E.M. Voorhees: *Overview of the TREC 2003 question answering track*. 12th Text REtrieval Conference (TREC 2003), pp. 54-68, Gaithersburg, 2004

[49] J.R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, M. Tyson: *FASTUS: A Cascaded Finite State Transducer for Extracting Information from Natural Language Text*. Finite State Devices for Natural Language Processing, MIT Press, Cambridge, 1996

[50] E. Riloff: *Automatically Constructing a Dictionary for Information Extraction Tasks*. 11th National Conference on Artificial Intelligence (AAAI-93), pp. 811-816, Washington, 1993

[51] J.T. Kim, D.I. Moldovan: *Acquisition of Semantic Patterns for Information Extraction from Corpora.* 9th Conference on Artificial Intelligence for Applications, pp. 171-176, Orlando, 1993

[52] C. Cardie: *Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis*. 11th National Conference on Artificial Intelligence (AAAI 1993), pp. 798-803, Washington, 1993

[53] P. Hastings, S. Lytinen: *The Ups and Downs of Lexical Acquisition*. 12th National Conference on Artificial Intelligence (AAAI 1994), pp. 754-759, Seattle, 1994

[54] R. Grishman: *Information extraction: Techniques and challenges*. Information Extraction: techniques and challenges, Vol 1299, pp.10-27, Springer, Berlin, 1997

[55] F. Ciravegna, A. Lavelli, G. Satta: *Bringing Information Extraction out of the Labs: the Pinocchio Environment*. 14th European Conference on Artificial Intellegence (ECAI 2000) on Machine learning for Information Extraction, pp. 416-420, Berlin, 2000

[56] R. Weischedel, S. Boisen, D. Bikel, R. Bobrow, M. Crystal, W.Ferguson: *Progress in information extraction*. workshop on held at Vienna, Association for Computational Linguistics, pp. 127-138, Vienna, 1996

[57] D. Kelly, J. Lin: *Overview of the TREC 2006 ciQA task*. ACM SIGIR Forum, Vol. 41, N. 1, pp. 107–116, ACM, New York, 2007

[58] C.A. Thompson, M.E. Califf, R. J. Mooney: *Active learning for natural language parsing and information extraction*. 16th International Conference on Machine Learning (ICMLA'99), pp. 406–414, Bled, 1999

[59] D. de Chalendar, T. Dalmas, F. Elkateb-Gara, O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, A. Vilnat: *The Question Answering System QALC at LIMSI: Experiments in Using Web and WordNet*. 11th Text Retrieval Conference (TREC 2002), Gaithersburg, 2002

[60] B. Popov, A. Kiryakov, D. Manov, A. Kirilov, D. Ognyanoff, M. Goranov: *Towards Semantic Web Information Extraction*. 2nd International Semantic Web Conference (ISWC2003), Sanibel Island, 2003

[61] S. Soderland: *Learning Information Extraction Rules for Semi-structured and Free Text*. Machine Learning, Vol. 34, pp. 233-272, Springer, Dordrecht, 1999

[62] A. Maedche, G. Neumann, S. Staab: *Bootstrapping an Ontology-Based Information Extraction System*. Intelligent Exploration of the Web, pp. 345-359, Springer, Berlin, 2003

[63] D. Freitag: *Machine Learning for Information Extraction in Informal Domains*. PhD thesis, Computer Science Department, Carnegie Mellon University, 1998

[64] E.M. Voorhees: *Natural Language Processing and Information Retrieval*. Lecture notes in computer science, Vol. 1714, pp. 32-48, Springer, Berlin, 1999

[65] M.T. Pazienza (editeur): *Information Extraction: a multidisciplinary approach to an emerging information technology*. Lecture notes in Computer Science, Vol. 1299, pp. 44-72, Springer, Berlin, 1997

[66] R. Gaizauskas, W. Yorick: *Information Extraction: Beyond Document Retrieval*. Journal of Documentation, Vol. 54, N. 1, pp. 70-105, Emerald, Bradford, 1998

[67] N.A. Chinchor: *Overview of MUC-7/MET-2*. 7th Message Understanding Conference (MUC-7), 1998

[68] J. R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Tyson: *FASTUS: a system for extracting information from text*. Workshop on Human Language Technology, pp. 133-137, Princeton, 1993

[69] N. Chinchor, *MUC-7 Named Entity Task Definition*. 1997 http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html,

[70] A. Kehler, D. Appelt, L. Taylor, A. Simma: *The (non) utility of predicate-argument frequencies for pronoun interpretation*. Human Language Technology Conference of the North American

Chapter of the Association for Computational Linguistics (HLT/NAACL), pp. 289-296, Boston, 2004

[71] V. Ng: *Shallow Semantics for Coreference Resolution*. 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 1689-1694, Hyderabad, 2007

[72] R. Mitkov, B. Boguraev, S. Lappin: *Introduction to the special issue on computational anaphora resolution*. Computational Linguistics, Vol. 27, N° 4, pp. 473–477, MIT Press Cambridge, 2001

[73] N. Chinchor: *MUC-7 Information Extraction Task Definition*. 1998 http://www.nlp.org.cn/docs/20030724/resource/Brief%20Definition%20of%20Information%20Extraction%20Task.htm

[74] B. Sundheim*: Overview of the results of the MUC-6 evaluation*. 6th Message Understanding Conference (MUC-6). Columbia, 1995

[75] M. Silberztein: *Dictionnaires électroniques et analyse automatique de textes* : *le système INTEX*. Masson, Paris, 1993

[76] Y. asaki, Y. Matsuo: *Learning Semantic-Level Information Extraction Rules by Type-Oriented ILP*. 18th International Conference on Computational Linguistics (COLING-2000), pp. 698-704, Saarbrücken, 2000

[77] F. Béchet, A. Nasr, F. Genet: *Tagging Unknown Proper Names Using Decision Trees*. 38th Annual Meeting of the Association for ComputationalLinguistics (ACL'2000), pp.77-84, Hong-Kong, 2000

[78] F. Kubala, R. Schwartz, R. Stone, R. Weischedel: *Named entity extraction from speech*. DARPA Broadcast News Transcription and Understanding Workshop, Herndon, 1998

[79] S. Sekine, Y. Eriguchi: *Japanese named entity extraction evaluation: analysis of results*. 18th conference on Computational linguistics (COLING-2000), pp. 1106-1110, Saarbrücken, 2000

[80] J.J. Burger, D. Day, L. Hirschman, P. Robinson, M. Vilain: *MITRE: Description of the Alembic System used for MUC-6*. 6th conference on Message understanding (MUC-6), pp. 141-155, Los Altos, 1995

[81] H. Tanev, B. Magnini: *Weakly Supervised Approaches for Ontology Population*. 1st Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, 2006

[82] P. Buitelaar, P. Cimiano, B. Magnini (éditeurs): *Ontology Learning from Text: Methods, Evaluation and Applications*. Vol. 125, IOS Press, Amsterdam, 2005

[83] P. Cimiano, A. Pivk, L.S. Thieme, S. Staab: *Learning Taxonomic Relations from Heterogeneous Sources of Evidence*. Ontology Learning from Text: Methods, Evaluation and Applications, pp. 59-73, IOS Press, Amsterdam, 2005

[84] B. Magnini, M. Negri, E. Pianta, L. Romano, M. Speranza, L. Serafini, C. Girardi, V. Bartalesi, R. Sprugnoli: *From Text to Knowledge for the Semantic Web: the ONTOTEXT Project*. Workshop on Semantic Web Applications and Perspectives (SWAP 2005), Trento, 2005

[85] M. Califf, R.J. Mooney: *Relational Learning of Pattern-Match Rules for Information Extraction*. 16th National Conference on Artificial Intelligence (AAAI-99), pp. 328-334 Orlando, 1999

[86] S.B. Huffman: *Learning information extraction patterns from examples*. Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, pp. 246-260, Springer, Berlin, 1996

[87] S. Ray, M. Craven: *Representing sentence structure in hidden Markov models for information extraction*. 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), pp. 1273-1279, Seattle, 2001

[88] *Méthode d'AIde à l'Innovation MAL'IN*. http://www.trefle.u-bordeaux1.fr/malin/index.html

[89] J.M. Zelle, R.J. Mooney: *Combining top-down and bottom-up methods in inductive logic programming*. 11th International Conference on Machine Learning (ML-94), pp. 343-351, New Brunswick, 1994

[90] A. Budanitsky, G. Hirst: *Evaluating wordnet-based measures of lexical semantic relatedness*. Computational Linguistics, Vol. 32, pp. 13-47, MIT Press, Cambridge, 2006

[91] R. Yangarber, R. Grishman, P. Tapanainen, S. Huttunen: *Unsupervised Discovery of Scenario-Level Patterns for Information Extraction*. 6th Conference on Applied Natural Language Processing (ANLP 2000), pp. 282-289, Seattle, 2000

[92] R. Yangarber, R. Grishman, P. Tapanainen, S. Huttunen: *Automatic Acquisiton of Domain Knowledge for information Extraction*. 18th International Conference on Computational Linguistics (COLING 2000), pp.940-946, 2000

[93] E. Riloff, R. Jones: *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping*. National Conference on Artificial Intelligence (AAAI-99), pp. 474-479, Orlando, 1999

[94] M.E. Califf, R.J. Mooney: *Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction*. Journal of Machine Learning Research, Vol. 4, pp. 177-210, MIT Press, Cambridge, 2003

[95] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates: *Web-scale Information extraction in KnowItAll (preliminary results)*. 13th International World Wide Web Conference (WWW2004), pp. 100-109, New York, 2004

[96] M.L. Reinberger, P. Spyns: *Unsupervised text mining for the learning of dogma-inspired ontologies*. Ontology Learning from Text: Methods, Applications and Evaluation, pp. 29-43, IOS Press, Amsterdam, 2005

[97] Z. Harris: *Mathematical Structures of Language*. John Wiley Interscience, New York, 1968

[98] S.Muggleton, C. Feng: *Efficient induction of logic programs*. 1st Conference on Algorithmic Learning Theory, pp. 368-381, Ohmsha, 1992

[99] S. Muggleton: *Inverse entailment and Progol*. New Generation Computing Journal, Vol. 13, pp. 245-286, Springer, Berlin, 1995

[100] D. Freitag, A.L. McCallum: *Information extraction using HMMs and shrinkage*. AAAI-99 Workshop on Machine Learning for Information Extraction, pp 31-36, Orlando, 1999

[101] L.A. Birnbaum, G.C. Collins (editeurs): *Learning Relations*. 8th International Workshop on Machine Learning, Part VI, Evanston, 1991

[102] J.R. Quinlan: *Learning logical definitions from relations*. Machine Learning, Vol. 5, N.3, pp. 239-266, Springer, Netherlands, 1990

[103] J. R. Quinlan: *Induction of decision trees*. Machine Learning, Vol. 1, N.1, pp. 81–106, Kluwer Academic Publishers, Hingham, 1986

[104] D. Fisher, S. Soderland, J. McCarthy, F. Feng, W. Lehnert: *Description of the UMass system as used for MUC-6*. 6th Message Understanding Conference (MUC-6), pp. 221-236, San Francisco, 1995

[105] University of Stuttgart, *TreeTagger: a language independent part-of-speech tagger*, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[106] Xerox Research Centre Europe (XRCE), http://www.xrce.xerox.com

[107] C. Leacock, C. Martin: *Combining local context and WordNet similarity for word sense identification*. WordNet: An Electronic Lexical Database, pp. 265–283, MIT Press, Cambridge, 1998

[108] I. Muslea: *Extraction Patterns for Information Extractions Tasks: A Survey*. AAAI'99 Workshop on Machine Learning for Information Extraction, Orlando,1999

[109] J. Liang, T. Nguyen, K. Koperski, G. Marchisio: *Ontology-Based Natural Language Query Processing for the Biological Domain*. Workshop on Linking Natural Language Processing and Biology (BioNLP at HLT-NAACL 06), pp. 9–16, New York, 2006

[110] G. MILLER: *Wordnet: A lexical database for English*. Communications of the ACM, Vol. 38, N. 11, pp. 39-41, Association for Computing Machinery, New York, 1995

[111] V.M. Tsourikov, L.S. Batchilo, I.V. Sovpel: *Document semantic analysis/ selection with knowledge creativity capability using subject-action-object (SAO) structures*. United States patent N° 6167370, 2000

[112] F. Ciravegna, A. Lavelli: *Full text parsing using cascades of rules: An information extraction perspective*. 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99), pp. 102-109, Bergen, 1999

[113] S. Cranefield: *UML and the Semantic Web*. International Semantic Web Working Symposium (SWWS2001), pp. 113-130, Palo Alto, 2001

[114] S. Cranefield, M. Purvis: *UML as an Ontology Modeling Language*. IJCAI-99 Workshop on Intelligent Information Integration, Stockholm, 1999

[115] S. Abiteboul, O.M. Duschka: *Complexity of answering queries using materialized views*. 17th ACM-SIGMOD symposium on Principles of database systems (PODS 1998), pp. 254-263, Seattle, 1998

[116] F. Baader, W. Nutt: *Basic Description Logics*. Description ogic Handbook, Theory, Implementation, and Applications, pp. 43-95, Cambridge University Press, New York, 2003

[117] N. Ide, J. Véronis: *Word sense disambiguation: The state of the art*. Computational Linguistics, Vol. 24, pp. 1-40, MIT Press, Cambridge, 1998

[118] D. Vickrey, L. Biewald, M. Teyssier, D. Koller: *Word-sense disambiguation for machine translation*. Conference on Empirical Methods in Natural Language Processing (EMNLP05), pp. 771-778, Vancouver, 2005

[119] D. Yarowsky: *Unsupervised word sense disambiguation rivaling supervised methods*. 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189–196, Cambridge, Massachusetts, 1995

[120] U.S. Kohomban, W.S. Lee: *Learning semantic classes for word sense disambiguation*. 43rd Annual Meeting of the Association for Computional Linguistics (ACL'05), Michigan, 2005

[121] C. Santamarìa, J. Gonzalo, F. Verdejo: *Automatic association of web directories with word senses.* Computacional Linguistics, Vol. 29, pp. 485-502, MIT Press, Cambridge, 2003

[122] *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation, 2004 http://www.w3.org/TR/rdf-schema/

[123] F. Baader: *Logic-based knowledge representation*. Artificial Intelligence Today, Recent Trends and Developments, Vol. 1600, pp. 13-41, Springer, Berlin, 1999

[124] Protégé. http://protege.stanford.edu/

[125] *Resource Description Framework (RDF)*. http://www.w3.org/RDF/

[126] *OWL Web Ontology Language: Overview*. W3C Recommendation, 2004 http://www.w3.org/TR/owl-features/

[127] H. Schmid: *Probabilistic Part-of-speech tagging using decision trees*. International Conference on New Methods in Language Processing (NeMLaP), Manchester, 1994

[128] I. Al haj Hasan, M. Schneider, G. Gogu: *Automatic Feeding of an Innovation Knowledge Base Using a Semantic Representation of Field Knowledge*. OTM Conferences, pp.1034-1049, Vilamoura, 2007

Annexe A

Principes inventifs

Le tableau ci-dessous représente les principes inventifs tels qu'ils sont définis dans la base de connaissance TRIZ.

N.	Principe	Significations
1	Segmentation	Diviser un objet en pièces indépendantes
		Faciliter le désassemblage d'un objet
		Accroître le degré de fragmentation d'un objet
2	Extraction	Extraire un élément ou une propriété nuisible d'un objet ou isoler l'élément ou la propriété utile d'un objet
3	Qualité locale	Faire passer la structure d'un objet, un environnement ou une influence externe d'un état homogène à un état hétérogène
		Amener chaque partie fonctionnelle de l'objet dans les conditions de fonctionnement appropriées
		Amener chaque partie de l'objet à remplir une fonction utile et différente.
4	Asymétrie	Remplacer la forme symétrique d'un objet par une forme asymétrique
		Si la forme est déjà asymétrique, accroître son degré d'asymétrie
5	Combinaison	Rapprocher ou fusionner des objets identiques ou similaires, assembler des parties identiques ou similaires pour réaliser des opérations parallèles
		Combiner ou paralléliser des actions, les rapprocher dans le temps
6	Universalité	Faire en sorte que l'objet assure plusieurs fonctions, de manière éliminer le besoin d'autres pièces
7	Placement intérieur « poupées russes »	Placer les objets en série les uns dans les autres
		Faire passer un élément dans une cavité d'un autre
8	Contrepoids	Contrebalancer la masse d'un objet en le combinant avec un autre qui le soulève.
		Compenser la masse d'un objet en le faisant interagir avec son environnement (par exemple, en utilisant une force aérodynamique, hydrodynamique etc.)
9	Action inverse préliminaire	Si l'action à exécuter présente à la fois des effets utiles et néfastes, celle-ci devra être précédée d'actions inverses contrôlant les effets néfastes.
		Créer des contraintes internes de l'objet, qui s'opposeront aux contraintes néfastes de l'objet en fonctionnement.

10	Action préliminaire	Réaliser à l'avance (entièrement ou partiellement) un changement requis plus tard
		Prépositionner idéalement les objets de façon à ce qu'ils entrent en action efficacement et sans perte de temps
11	Compensation ou protection préliminaire	Compenser le manque de fiabilité de l'objet par des contremesures prises à l'avance
12	Equipotentialité	Limiter les changements de position (par exemple, changer les conditions de travail de manière à ce que l'objet n'ait besoin ni d'être élevé ni d'être abaissé)
13	Inversion	Inverser l'action utilisée pour résoudre le problème (par exemple, au lieu de refroidir un objet, le réchauffer)
		Rendre les pièces mobiles fixes et inversement
		Retourner l'objet (ou le procédé)
14	Sphéricité	Remplacer des parties, surfaces ou formes rectilignes par des curvilignes, des surfaces planes par des surfaces sphériques ou des pièces parallélépipédiques par des structures sphériques
		Utiliser des rouleaux, des billes, des spirales, des dômes
		Remplacer une translation par une rotation ; utiliser la force centrifuge
15	Mobilité	Permettre ou concevoir une optimisation des caractéristiques de l'objet, de l'environnement extérieur ou du procédé ou trouver des conditions de fonctionnement optimales
		Diviser un objet en plusieurs éléments mobiles les uns par rapport aux autres
		Si un objet (ou un procédé) est fixe, le rendre mobile ou adaptable
16	Action partielle ou excessive	S'il est difficile d'obtenir un effet à 100% par une méthode donnée, appliquer cette méthode « partiellement » ou « à l'excès » peut simplifier considérablement le problème
17	Changement de dimension	Déplacer un objet dans un espace bidimensionnel ou Tridimensionnel
		Utiliser un assemblage multicouche d'objets plutôt qu'un assemblage monocouche
		Incliner ou réorienter l'objet, le poser de côté
		Utiliser l'autre face d'une surface donnée
18	Vibration mécanique	Faire osciller ou vibrer un objet
		Si l'oscillation existe, accroître sa fréquence
		Utiliser la fréquence de résonance d'un objet
		Utiliser des vibrations piézo-électriques au lieu de mécanique
		Combiner ultrasons et champ électromagnétique
19	Action périodique	Remplacer une action continue par une action périodique ou pulsatoire
		Si l'action est déjà périodique, modifier sa fréquence ou son amplitude

		Utiliser les pauses entre les impulsions pour accomplir une autre action
20	Continuité d'une action utile	Privilégier une action continue (sans pauses), où toutes les parties d'un objet agissent plein régime
		Eliminer tous les temps morts
21	Grande vitesse	Effectuer un procédé ou certaines phases dangereuses ou néfastes à grande vitesse
22	Application bénéfique d'un effet néfaste	Utiliser des facteurs néfastes (en particulier les effets néfastes de l'environnement) pour obtenir un effet positif
		Annuler l'effet d'un facteur néfaste en le combinant avec un autre facteur néfaste
		Accroître un effet néfaste jusqu'à ce qu'il ne soit plus nuisible
23	Asservissement	Introduire un asservissement (réponse, vérification) afin d'améliorer un procédé ou une action
		Si l'asservissement existe déjà, modifier son ampleur ou son influence
24	Intermédiaire	Utiliser un objet ou un procédé intermédiaire
		Combiner provisoirement un objet à un autre (opération facilement réversible)
25	Self-service	Faire de sorte que l'objet se suffise à lui-même en effectuant des fonctions auxiliaires utiles
		Réutiliser les résidus énergétiques et matériels
26	Copie	Utiliser des copies simplifiées et bon marché plutôt qu'un objet complexe, cher ou fragile
		Remplacer un objet ou un procédé par sa copie optique Si les copies optiques sont déjà utilisées, utiliser les copies infrarouges ou ultraviolettes
27	Ephémère et bon marché	Remplacer un objet cher par un ensemble d'autres objets bon marché, en renonçant à certaines de ses qualités (comme la durée de l'action par exemple)
28	Remplacer les Systèmes mécaniques	Remplacer un système mécanique par un système sensoriel (optique, acoustique, olfactif)
		Utiliser des champs électriques, magnétiques, électromagnétiques pour interagir avec l'objet
		Remplacer les champs statiques par des champs mobiles, les champs aléatoires par des champs structurés
		Utiliser les champs en conjonction avec des particules activées par ces champs (par exemple, ferromagnétiques)
29	Systèmes pneumatiques et hydrauliques	Remplacer les parties solides d'un objet par un gaz ou un liquide ; par exemple, objets gonflables (à air ou eau), à coussin d'air, hydrostatique ou hydro-réactif
30	30 Membrane flexible et film mince	Remplacer les structures tridimensionnelles par des membranes flexibles et des films minces
		Isoler l'objet de son environnement en utilisant des membranes flexibles et des films minces

31	Matériau poreux	Rendre un objet poreux ou lui adjoindre des éléments poreux
		Si l'objet est déjà poreux, remplir les pores d'une substance ou d'une fonction utile
32	Changement de couleur	Modifier la couleur d'un objet ou de son environnement extérieur
		Modifier le degré de transparence d'un objet ou de son environnement extérieur
33	Homogénéité	Faire interagir les objets avec un objet annexe de même matière (ou d'une matière ayant des propriétés identiques)
34	Eliminer récupérer	Eliminer un élément de l'objet (par dissolution, évaporation etc.) lorsque celui-ci a assuré sa fonction ou le modifier au cours de fonctionnement
		A l'inverse, récupérer les éléments consommables de l'objet au cours du fonctionnement
35	Changement de paramètre	Modifier l'état physique d'un objet (ex. sous forme de gaz, de liquide ou de solide)
		Modifier l'état physique d'un objet (ex. sous forme de gaz, de liquide ou de solide)
		Modifier le degré de flexibilité
		Modifier la température
36	Changement de phase	Utiliser les phénomènes associés aux changements de phase (changement de volume, perte ou absorption de chaleur etc.)
37	Dilatation thermique	Utiliser la dilatation ou la contraction thermique des matériaux
		Utiliser des matériaux différents avec des coefficients de dilatation différents
38	38 Oxydants puissants	Remplacer de l'air normal par de l'air enrichi
		Remplacer l'air ou l'oxygène par des radiations ionisantes
		Utiliser de l'oxygène ozonisé
		Remplacer l'oxygène ozonisé ou ionisé par de l'ozone
39	Environnement inerte	Remplacer un environnement normal par un environnement inerte
		Ajouter des pièces neutres ou des additifs inertes à un objet

Annexe B

Extrait des effets scientifiques définis dans TechOptimizer

Function	Function
Abrupt reverse-biased p-n junction	Ionic polarization of dielectric
Absorbed radiation energy heats body	Ionic strength of solution affects acid dissociation constant
Absorption (absorption of gas by liquid)	Ionic-adsorption field effect
Absorption (dissolution of gas in liquid)	Ionic-ionic emission
Absorption (separation of gas mixture)	Ionization
Absorption factor dependence on wavelength	Ionization losses of charged particles
Absorption layer thickness determines light intensity	Ionized cluster beam deposition
Absorption of light by aerosol	Ionized gas charges aerosol particles
Absorption of microwave radiation by diatomic molecule	Ionizing radiation produces hydrogen peroxide
Absorption of microwave radiation during passing through moist materials	Ionizing shock wave creates potential difference
Absorption of sound in substance affects sound reverberation	Ionoluminescence
Absorption of X-rays	Ions charge solid particles
Absorption/desorption of hydrogen in intermetallic compounds	Ions passing through crystal increase their charge
Ac magnetic field changes voltage across coil	Irradiating substance with ions produces new substance
Ac magnetic field creates shear wave	Irreversible dissolution (etching)
AC signal demodulation by means of Josephson junction	Isothermicity of heat pipe condenser surface
Accelerated electrons produce ozone from air molecules	Isotope indication
Accelerated motion of body	Isotropic wet chemical etching
Accelerated rotation of body	Jet attachment (Coanda effect)
Accelerated steam flow increases heat transfer coefficient	Jet erosion
Acceleration of conductor generates electric current	Jet exhaust velocity determines length of gas jet
Acceleration of object changes its velocity	Jet expands at capillary outlet
Acceleration of solid	Jet induces flow in stationary fluid
Accumulation of electric energy in capacitor	Jet pressure changes angle of jet separation from surface
Accumulation of thermal energy by	Jet speed changes heat transfer coefficient

changing temperature of substance	
…	…
…	…
Elastic medium density changes transverse wave speed	Refraction of light
Elastic modulus of concentrated electrorheological suspensions	Refractive index of medium affects light intensity
Elastic scattering of electrons	Refractive index variation
Elastic-plastic deformation	Regeneration of catalyst in fluidized bed
Elasticity modulus affects normal compressive stress	Regulation of solid particle mixing in fluidized bed
Elasticity modulus affects normal tensile stress	Relation between electroluminescence intensity and direct current
Elasticity of magnetic fluid film	Relation between photodetector sensitivity and concentration gradient
Electric conductivity of blood	Relation between photodetector sensitivity and p-n junction depth
Electric current controls electrodeposition rate	Relationship between photodiode breakdown voltage and bandgap
Electric current destroys plasma	Relative motion of solids determines friction between them
Electric current destroys superconducting state	Relative permeability determines reluctance
Electric current from radioisotope decay	Relaxation time determines amplitude of electron spin echo
Electric current heats electrolyte	Relaxation time determines amplitude of nuclear spin echo
Electric current initiates lasing in p-n junction	Relaxation time sets longitudinal nuclear magnetization
Electric current removes microparticles	Reluctance sets magnetic flux
Electric current rolls cylinder along plain support	Removal of substance by mechanical action
Electric discharge creates nanotubes	Reservoir volume defines gas pressure change
Electric discharge in gas	Residual magnetization of ferromagnet
Electric discharge lowers nitrogen monoxide concentration	Resistance affected by conducting liquid
Electric field affects density of pseudo-liquefied layer	Resistance of conductor changes thermal power
Electric field affects drift velocity of ions	Resistance of superconductor tunnel junction
Electric field affects electron temperature of plasma	Resistance-composition relation
Electric field amplifies sound in piezosemiconductor	Resistive strip length determines output voltage
…	…
…	…
Grain size affects polycrystal yield strength	Temperature stratification in liquid
Granular material drying in fluidized bed	Tension-compression of solids (Poisson's effect)
Granulation in fluidized bed	Tensoconductance of metal-filled polymers
Graphoepitaxy	Tensoresistive effect I
Gravitation	Tensoresistive effect II

Gravitation changes refractive index of neutrons	Thermal action produced by Foucault currents
Gravitational interaction	Thermal autoelectronic emission
Gravitational moment	Thermal breakdown of dielectric
Gravitational pressure gradient	Thermal Christiansen effect
Gravitational surface wave length affects wave velocity	Thermal conductance of metal-filled polymers
Gravity affects vaporization of liquid	Thermal conductivity determines time to heat surface
Gregorig effect	Thermal contact resistance
Griffith's effect	Thermal decomposition
Grounding	Thermal decomposition of metal carbonyls
Growth rate determines quantity of animal cells in culture	Thermal effect of hydrogen absorption/desorption in intermetallic compounds
Guest-host effect in nematic liquid crystals (NLC)	Thermal electron emission (Richardson effect)
Gunn effect	Thermal expansion of bi-substances
Gunn oscillation frequency switching by semiconductor shape	Thermal expansion of solid bodies
Gyroscope precession	Thermal lens effect
Gyroscopic effect	Thermal oxidation of silicon at atmospheric pressure
Gyrotropic deflection of magnetic bubbles in gradient bias field	Thermal radiation dependence on surface roughness
Hall effect	Thermal resistance reduction of contacting bodies utilizing liquid filler
Hartmann generator	Thermal-assisted external photoemission
Head-on collisions with electrons increase photon frequency	Thermo-optical effect in smectic liquid crystals (SLC)
Heat accumulation	Thermoacoustic effect
Heat accumulation during substance melting	Thermocapillary effect
Heat and electric conduction in heat pipe	Thermochemical processing of metals in fluidized bed
Heat capacity affects duration of heating object surface	Thermochromic effect in cholesteric liquid crystals (CLC)
Heat conduction (creation of temperature difference)	Thermochromism
Heat conduction of rarefied gas (influence of pressure)	Thermoconvective instability effect
Heat conduction of solids temperature dependence	Thermodestruction of polymers
Heat conductivity change in gas due to pressure	Thermodiffusion
Heat conductivity effect on temperature gradient	Thermoelectromotive force (Seebeck effect)
Heat conversion effect of heat pipe	Thermoferromagnetic effect
Heat deforms solid	Thermoionic emission
Heat determines plasma magnetization	Thermomagnetic convection effect
Heat diode effect in gravity-assisted heat pipe	Thermomagnetic convective instability
Heat exchange dependence on velocity	Thermomagnetic effect
Heat exchange from solid to fluid	Thermomagnetic effect of magnetic fluid

Heat flow affects diffusion of impurity in fluid	Thermomechanical effect
Heat flow dependence on conductivity	Thermonuclear fusion heating
…	…
…	…
Infrared light amplifies luminescent radiation	Viscosity temperature dependence effect
Infrared radiation destroys protein	Viscous liquid decelerates droplets of another liquid
Initiation of gamma-radiation	Viscous liquid decelerates gas bubbles
Initiator efficiency affects polymerization rate	Voltage creation by pyroelectric effect
Injection electroluminescence	Voltage transformer
Injection from wide-band emitter	Voltage-modulated photon-assisted tunneling
Insulator influencing threshold voltage of field transistor	Volume change affects absorbed power of magnetic field
Intensity of optical beam affects its filamentary structure	Volume vapor condensation
Inter-mode interference change from axial strain on fiber optic light guide	Vortex rotates fluid
Interference fringes' period dependence on angle between interfering waves	Vortices penetration into Josephson junction
Interference of light	Wannier-Mott excitons
Intermetallic	Water density anomalous temperature dependence
Internal photoemission threshold in heterostructure	Water flow creates ice layer
Internal photoresistive effect	Water hydrolyzes substance
Internal wave stirs liquid	Water jet range increase using polymer additions
Intrinsic absorption of light	Wavelength affects damping of microwave in waveguide
Intrinsic photoeffect threshold	Wavelength affects frequency of spin wave
Inverse Wiedemann effect	Wavelength dependence of ellipticity of light
Inversion layer formation	Wavelength determines concentration of radiant energy density
Ion assisted deposition	Wavelength dispersion of microbending losses
Ion beam produces stress in substance	Wavelength of light affects interference pattern intensity
Ion beam sputtering	Weak electrolyte concentration controls dissociation degree
Ion charge affects drift velocity of ions	Wedge
Ion concentration affects interaction energy of plates	Weissenberg effect
Ion conductivity of solutions	Wetting
Ion current density determines ion implantation rate	Wetting angle determines capillary filling speed
Ion currents in gases	Wetting angle determines surface curvature radius
Ion exchange	Wetting angle of capillary determines fluid height
Ion exchange with ionites	Wheel diameter affects energy required to move solid
Ion implantation	Wiegand effect
Ion irradiation produces amorphous substances	Wing bursts vortices
Ion mass affects mobility of ions	Worm pair

Ion radii affects lattice energy	X-ray diffraction
Ion strength sets equilibrium constant of complex ion stability	X-ray lithography
Ion strength sets equilibrium constant of substance solubility	X-ray total external reflection
Ion temperature affects radiation spectrum of plasma	X-rays create sound waves in metal
Ion-beam lithography	X-rays produce oximes in organic acid solution
Ion-electron amplification	Zeeman effect
Ion-implantation doping through semi-permeable mask	Zenith angle of Sun affects altitude of Earth's ionosphere
Ion-optical amplification	Zeolite absorbs water vapor
Ion-photon emission	Zone melting (thermogradient zone migration)

Annexe C

Ressources

Les ressources définies dans TechOptimizer et utilisées pour formuler les effets ou les solutions innovantes génériques sont les suivantes.

Resource	**Resource**	**Resource**
Abrasive suspension	Crack	Fire
Acid	Crystal	Flow
Acoustic wave	Current	Flower
Adhesive substance	Cutting force	Fluid
Aerosol	Damper	Fluidic prism
Air	Damping elements	Fluorescent material
Air-fuel mixture	Detergent	Flywheel
Alternating magnetic field	Diaphragm	Force
Alternating voltage	Dichroïc plate	Force field
Ampere force	Dielectric crystal	Friction
Anhydrous salt	Dielectric film	Gamma radiation
Anisotropic medium	Dielectric liquid	Gas
Antenna	Dielectric particles	Gem
Arc discharge	Dielectric substance	Granular material
Avalanche discharge	Diode	Gravity
Balanced weight	Disbalanced mass	Grid
Ball	Dopant	Gunn diode
Barkhausen noise	Doped glass	Gyroscopic effect
Battery	Double-paned window	Hall sensor
Beam	Earthed plate	Heat
Beam splitter	Elastic cushion	Heavy current device
Birefringent plate	Elastic material	Holographic material
Body	Elastomeric member	Hot air
Brush	Electret	Hydrazine
Buoyant force	Electric Field	Hyperboloïd reflector
Capacitor	Electric probe	Infrared light
Catalyst	Electrochromic film	Interference pattern
Cavity	Electrode	Ion
Centrifugal force	Electrolyte	Ionic Solution
Chemical material	Electromagnet	Isotope
Coil	Electromagnetic field	Kerr effect filter
Collector	Electromotive force	Knurl
Combined dielectric material	Electron	Laser beam
Compressed air	Electron beam	Lens

Compressed gas	Electro-optic liquid	Light
Concentrator	Ellipsometer	Liquid
Condenser	Emitter	Liquid air
Conductor	Ferro magnet	Liquid crystal
Cone	Ferroelectric material	Liquid jet
Contacted conductor	Ferromagnetic	Load
Contacting conductor	Ferromagnetic core	Loose material
Coolant	Ferromagnetic liquid	Lorentz force
Cooling liquid	Ferromagnetic particles	Magnet
Core	Filler	Magnetic field
Corneal membrane	Film	Magnetic fluid
Cotton-Mouton converter	Filter	Magnetic material
Magnetic sensor	Photodiode	Solid waste
Magneto hydrodynamic generator	Photoelastic material	Solution
Magnetooptical	Photon	Solvent
Magnetostrictive material	Photoresistor	Sound
Mass	Photovoltaic cell	Sound absorbing material
Material	Piezoelectric material	Substance
Medium	P-junction	Substrate
Melting substance	Plasma	Sulfur dioxide
Membrane	Plastic material	Superconductor
Metallic material	P-n junction	Surfactant
Metallic particles	Polarizer	Tachometer
Metal-plated strip	Polymer	Tangential stress
Microwave radiation	Porous material	Temperature
Mirror	Powder	Temperature gradient
Mixture	Pressure	Tension
Molten metal	Prism	Thermo electromotive material
Motion	Protector	Thermo sensitive material
Moving electric charge	Pyroelectric material	Thermocouple
Multi-layer structure	Quenching agent	Titanium coating
N-junction	Radiation	Torque
Nonlinear crystal	Receiver	Transducer
Non-uniformly magnetized element	Rectilinear conductor	Tunnel junction
Normal stress	Reflecting surface	Ultrasound
Optical fiber	Resonator	Ultraviolet
Optical isolator	Rotation	Unbalanced mass
Optical method	Semiconductor	Vacuum
Optical radiation	Sensor	Variable magnetic field
Optical sensor	Shape memory alloy	Vibration
Optical switch	Sharpened electrode	Vibratory system
Oscillation	Shell	Voltage
Oxygen	Silicon	Vortex tube
Palm surface	Silicon containing gas	Wave
Paraboloïd reflector	Slit	Wiegand material

Paramagnetic material	Solar radiation	Wiper
Particles	Photo conducting material	Zinc sulfide layer
Passivation layer	Photo resist layer	
Pendulum	Photochromic material	

Annexe D
Opérateurs et exemples

Dans cette annexe nous présentons des opérateurs d'innovation et des exemples de la base de connaissances de TechOptimiser. Ces opérateurs sont insérés dans notre base de connaissances pour mener nos expérimentations (cf. chapitres 4 et 5 et annexe F). Nous avons souligné dans les opérateurs les ressources et dans les exemples les objets des opérateurs.

Tout d'abord, quatre effets sont présentés : évaporation de liquide en diminuant la pression de sa vapeur saturée, pression à partir de l'impact liquide-solide, évaporation, fragmentation de liquide par cavitation, perdre de poids des objets immergés dans un liquide. Dans ces effets, les ressources utilisées sont des substances (*gaz*, *liquide*, *fluide*), des actions (*shock-wave*, *fragmente*, *pression*) et des paramètres (*poids*).

- Dans le premier effet, la pression sur un liquide est utilisée pour produire un *stock-wave*.
- Dans le deuxième effet, la *pression* est utilisée pour *vaporiser* un *fluide*.
- Dans le troisième effet, un *gaz* est utilisé pour *fragmenter* un *liquide*.
- Dans le quatrième effet, un *liquide* est utilisé pour *diminuer* le *poids*.

Les exemples associés à ces effets présentent des problèmes résolus par ces effets. Dans ces exemples, il faut repérer les ressources objets : rocher, matériau solide dans le premier exemple, *tret-butanol*, *butyl hydroperoxid* dans le deuxième, aérosol dans le troisième et billets, poudre compacte dans le quatrième exemple.

On trouve ensuite deux exemples de solution innovante générique qui utilisent une segmentation pour résoudre le problème d'innovation. Dans le premier exemple, une segmentation en poudre (*ferromagnetic powder*) est utilisée dans une valve pour contrôler le gaz (*gas flow*) entre des parties (*parts*) mobiles. Dans le deuxième exemple, une segmentation

en plusieurs parties (*three-layered panels*) est utilisée afin de supprimer le bruit et la vibration dans une automobile.

Enfin, deux exemples de résolution de contradictions entre des paramètres par le principe de segmentation sont présentés. Le premier exemple résout la contradiction entre la forme et la facilité de réparation dans un *excavator dipper* et le deuxième exemple la résout entre la surface d'un objet fixe et la température ou la pression dans une couche semi-conductrice.

A partir des opérateurs ainsi présentés, on remarque que

1. Une ressource (i.e. liquide, pressure, gaz) peut être utilisée pour résoudre plusieurs problèmes.

2. La difficulté du repérage des objets des opérateurs (cf. termes soulignés dans les exemples) dans les exemples par une approche linguistique. Il existe une diversité des rôles syntaxiques joués par ces termes dans le contexte des ressources des opérateurs.

3. Les opérateurs d'innovation ont des types et des natures (physiques, mathématique, chimique) différents ainsi que leurs ressources.

4. Les relations sémantiques terme-terme, terme-ressource et ressource-ressource sont importantes pour le repérage des termes correspondant à ces ressources dans les testes.

    - La relation terme-terme la plus importante est la relation d'hyperonymie.

    - Les relations ressources-ressources les plus intéressantes sont les relations définies par les opérateurs.

        - Les effets définissent une relation cause-conséquence entres les ressources d'innovation. par exemple dans le premier effet *shock-wave* est causé par une pression.

        - La solution innovante générique représente une relation ressource-états ; dans les deux exemples de segmentation présentés des états (partie, poudre) sont associés à la segmentation.

- Les principes définissent une relation de contradiction entre attributs.

Type	Opérateur
Effet	**Pressure from solid-liquid impact** When a solid strikes against a liquid, a shock wave forms. At the moment of impact, a very thin layer of liquid forms in which the velocity, pressure, and density are high. This compressed liquid layer is the shock wave. The compressed layer is followed by a decompressed layer with a lower pressure. The greater the sound intensity, the greater the shock wave velocity. If the velocity of the solid is less than the sound velocity, the shock wave detaches from the solid and propagates at a supersonic velocity upstream of it. Otherwise, the shock wave travels together with the solid. The wave takes the form of a cone, called the Mach cone.
Exemple	**Solid fragmentation by shock wave** The device is designed to fragment rock and other hard or impact material. It includes a water chamber that can store water under high pressure. The chamber outlet is connected to a straight tube with a free end section. The end is closed with a rupture disk. The tube can withstand the same water pressure as the chamber. An elongated, blind hole is drilled in the material to be fragmented. Its diameter only slightly exceeds the outside diameter of the end of the tube. Then the tube end is inserted into the hole. When the water pressure is high enough, the disk ruptures. As a result, a pulse of water with the peak pressure as great as 0.6×10^{12} Pascal and a rise time of about one millisecond is directed into the hole. The produced shock wave breaks the material.
Effet	**Vaporization of fluid under decrease in pressure of its saturated vapor** This effect occurs as a result of an interaction between the vapor and the fluid phases of the heat transfer agent. Under equilibrium, the saturated vapor pressure is relater to the temperature of the fluid with which it is in contact. A change in the saturated vapor pressure leads to adiabatic (without an external heat supply) fluid vaporization, accompanied by the fluid cooling to the vapour temperature. This process has a high rate because the latent heat of fluid vaporization is much higher than the heat capacity.
Exemple	**Decreasing pressure recovers tret-butanol from mixture** A pressure decrease is used to isolate tret-butanol. The reaction mixture (tret-butanol and butyl hydroperoxide) is placed in vacuum distillation device. The pressure above the mixture is reduced. The pressure of the saturated tret-butanol vapor decreases. Adiabatic vaporization of tret-butanol takes

	place as the saturated vapor pressure decreases. Tret-butanol passes from the reaction mixture into a vapor where it is recovered in its vaporization phase.
Effet	**Cavitation fragmentation of liquid** As the local pressure in the liquid decreases to the value close to the pressure of the liquid vapor, cavitation occurs producing a large quantity of bubbles filled with gas and vapor dissolved in the liquid. Cavitation cavities (vapor and gas bubbles) formed in the liquid flow result in rupture of continuity thereof and fragmentation.
Exemple	**Apparatus and method for producing liquid aerosol** To produce liquid aerosols a pre-dispersed liquid is fed onto the vibrating surface. Cavitation occurs in droplets leading to fragmentation thereof, and a high density fountain of aerosol particles is formed. Forces causing formation and collapse of cavities by vibrational cavitation are continuous high frequency pressure, oscillations with large amplitude. Cavities are formed when pulsation amplitude is sufficiently large, and liquid pressure drops to that of the saturated vapor or lower.
Effet	**Loss in weight of object immersed in fluid** Body weight compensation occurs due to buoyancy. The hydro-weightlessness effect is observed when the object is immersed in a fluid and its average density coincides with the object density. The weightlessness effect is exhibited for the system as a whole rather than for its internal structures. For example, the internal organs of a man lying motionless in the water may create the feeling of weightlessness. However, no changes in the fluid metabolism take place similar to those produced by true zero-gravity.
Exemple	**Density monitoring** Producing billets from the compacted powder with a preselected green density (before sintering) could provide high billet uniformity and product strength. The density monitoring proposed in the given invention makes it possible to determine a correct time of completion of charging of powder and the depressurization mode.

Solution innovante générique	**Segmentation** ➢ Diviser un objet en pièces indépendantes ➢ Faciliter le désassemblage d'un objet ➢ Accroître le degré de fragmentation d'un objet
Exemple	**Substance segmentation into powder** A magnet-controlled throttle valve closes an air passage. The magnetic field is generated by a solenoid (a coil of wire with a current flowing in it). **Disadvantage**: the design is unreliable since it does not assure a tight fit between the movable parts. **Solution**: it is proposed to make the valve using ferromagnetic powder placed between grates in the air passage. When acted on by a magnetic field, the powder aligns with the field lines and becomes compacted. This increases the pneumatic resistance. The magnetic field intensity controls the pneumatic resistance, enabling gas flow to be controlled.
Exemple	**Substance segmentation into several parts** Automobile bodies are fabricated from solid metal sheets **Disadvantage**: the automobile body panels *poorly damp noise and vibrations*. **Solution:** It is proposed to make three-layered panels using two metal sheets and a plastic insert. The sheets are joined by cement or spot welding. Such panels *are good noise and vibration suppressors*
Principe inventif	**Segmentation**
Exemple	**Segmenting a dipper lip** An excavator dipper has a lip formed as one piece of hard steel. If only a portion of it has been worn out or damaged the entire lip must be replaced. **Disadvantage**: This is a labour and time consuming job, causing the excavator to stand idle. **Solution:** It is proposed to use the segmentation principle to make the dipper lip more serviceable. One can segment the lip by forming it of separate detachable sections. This allows only the damaged or worn sections to be replaced quickly and easily.
Exemple	**Grooves prevent strain** IC's contain semi-conducting layers that form various components.

	Disadvantage: These components in operation can produce thermal strains between each other. **Solution:** It is proposed to use the segmentation principle to reduce the strains. The upper layers are divided into separated sections with grooves etched between those components that cause heating. The grooves inhibit the development of thermal strains due to differential heating, allowing each to expand independently.

Annexe E

Evaluation de l'extraction de triplets (sujet, action, objet) par des analyseurs syntaxiques

Nous avons expérimenté l'extraction de triplets avec quatre analyseurs syntaxiques : *Connexor Machinese Syntax*[14], *Link Grammar*[15], *XIP Parser*[16] et *Proximity Technolgy Parser*[17]. Tous ces analyseurs sauf *Link Grammar* sont des outils commerciaux. Nous les avons testés avec des versions de démonstration téléchargées depuis Internet et installées sur une machine locale. Pour cette expérimentation, nous nous sommes intéressés à la performance plus qu'à la technique de l'analyse. Ces analyseurs sont considérés comme étant les plus performants surtout les deux premiers. Ils sont utilisés dans plusieurs applications linguistiques (extraction d'information, traduction automatique, texte mining, désambiguïsation de sens…).

Dans cette annexe, nous présentons le limites des ces analyseurs par rapport à notre objectif. Cet objectif, comme présenté dans le chapitre 3, consiste à extraire des triplets (sujet, action, objet) pour permettre une comparaison avec une représentation similaire des opérateurs d'innovation.

Notre évaluation est basée sur les analyses syntaxiques retournées pour les deux phrases suivantes :

I. *Particles of the machined material have the same charge as the abrasive wheel and are repelled off its surface.*

II. *Particles of the machined material which have the same charge as the abrasive wheel are repelled off its surface.*

---

[14] Démonstration est sur le site http://www.connexor.eu/technology/machinese/demo/syntax/
[15] Téléchargement et démonstration sont sur le site http://www.link.cs.cmu.edu/link/
[16] Démonstration est sur http://www.xrce.xerox.com/xip/page1.jsp
[17] Démonstration est sur le site http://www.clres.com/

Ces phrases représentent des structures typiques rencontrées dans les documents techniques. Un analyseur syntaxique adapté à notre problème doit être capable d'identifier dans ces phrases les deux triplets suivants.

1. ("*particles of the machined material*", "*have*", "*charge*")
2. ("*particles of the machined material*", "*are repelled off*", "*its surface*")

Le premier triplet représente une fonction de rechargement par une charge électrique et le deuxième représente un effet de cette fonction qui correspond à une répulsion des objets chargés.

Dans la première phrase, ces deux triplets sont définis par deux clauses en conjonction. Par contre, dans la deuxième phrase, le premier triplet est introduit par une clause imbriquée dans le sujet de la phrase principale.

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	particles	particle	subj:>6	@SUBJ %NH N NOM PL
2	of	of	mod:>1	@<NOM-OF %N< PREP
3	the	the	det:>5	@DN> %>N DET
4	machined	machined	attr:>5	@A> %>N A ABS
5	material	material	pcomp:>2	@<P %NH N NOM SG
6	have	have	main:>0	@+FMAINV %VA V PRES
7	the	the	det:>9	@DN> %>N DET
8	same	same	attr:>9	@A> %>N A ABS
9	charge	charge	obj:>6	@OBJ %NH N NOM SG
10	as	as	mod:>9	@<NOM %N< PREP
11	the	the	det:>13	@DN> %>N DET
12	abrasive	abrasive	attr:>13	@A> %>N A ABS
13	wheel	wheel	pcomp:>10	@<P %NH N NOM SG
14	and	and	cc:>6	@CC %CC CC
15	are	be	v-ch:>16	@+FAUXV %AUX V PRES
16	repelled	repel	cc:>6	@-FMAINV %VP EN
17	off	off	ha:>16	@ADVL %EH PREP
18	its	it	attr:>19	@A> %>N PRON GEN SG3
19	surface	surface	pcomp:>17	@<P %NH N NOM SG
20	.	.		

Figure E.1: Analyse de la phrase I par *Connexor Machinese Syntax*

Les résultats des analyses, présentés dans les figures de E.1 à E.8, montrent que les quatre outils ont des difficultés pour bien identifier les sujets, les actions et les objets.

Dans la figure E.1, on peut considérer que dans l'analyse de la phrase I *Connexor Machinese Syntax* n'a pas commis d'erreur. Le résultat obtenu est suffisant pour extraire les triplets. Les sujets sont repérés par @subj dans la quatrième colonne, les verbes par %VA, %VP, %AUX dans la même colonne et leurs objets directs par @OBJ dans la troisième colonne. Les objets indirects dans la phrase n'ont pas un étiquetage spécifique mais on peut les extraire de l'étiquetage (pcomp) dans la troisième colonne. Les relations nécessaires pour identifier les éléments de chaque triplet peuvent être inférées par les chiffres qui mettent en relation les lignes des tableaux.

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	particles	particle		@SUBJ %NH N NOM PL
2	of	of	mod:>1	@<NOM-OF %N< PREP
3	the	the	det:>5	@DN> %>N DET
4	machined	machined	attr:>5	@A> %>N A ABS
5	material	material	pcomp:>2	@<P %NH N NOM SG
6	which	which	subj:>7	@SUBJ %NH <Rel> PRON WH NOM
7	have	have	mod:>5	@+FMAINV %VA V PRES
8	the	the	det:>10	@DN> %>N DET
9	same	same	attr:>10	@A> %>N A ABS
10	charge	charge	obj:>7	@OBJ %NH N NOM SG
11	as	as	pm:>16	@CS %CS CS
12	the	the	det:>14	@DN> %>N DET
13	abrasive	abrasive	attr:>14	@A> %>N A ABS
14	wheel	wheel	subj:>15	@SUBJ %NH N NOM SG
15	are	be	v-ch:>16	@+FAUXV %AUX V PRES
16	repelled	repel	man:>7	@-FMAINV %VP EN
17	off	off	ha:>16	@ADVL %EH PREP
18	its	it	attr:>19	@A> %>N PRON GEN SG3
19	surface	surface	pcomp:>17	@<P %NH N NOM SG
20	.	.		
21	<p>	<p>		

Figure E.2: Analyse de la phrase II par *Connexor Machinese Syntax*

Par contre, à cause de la clause *which*, l'analyseur n'a pas pu effectuer une analyse correcte de la phrase II (cf. figure E.2). Le mot *particles* est identifié comme sujet mais il n'est pas assigné à un verbe ou à un objet. Le mot *wheel* qui est complément dans la clause *which* est annoté comme sujet dans la phrase principale.

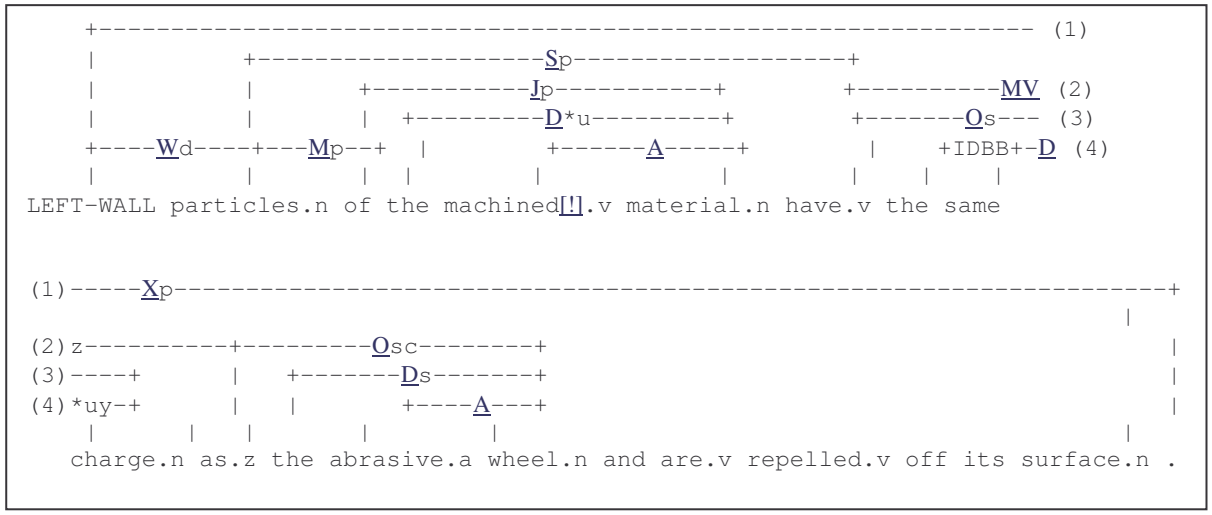


Figure E.3: Analyse de la phrase I par *Link Grammar*

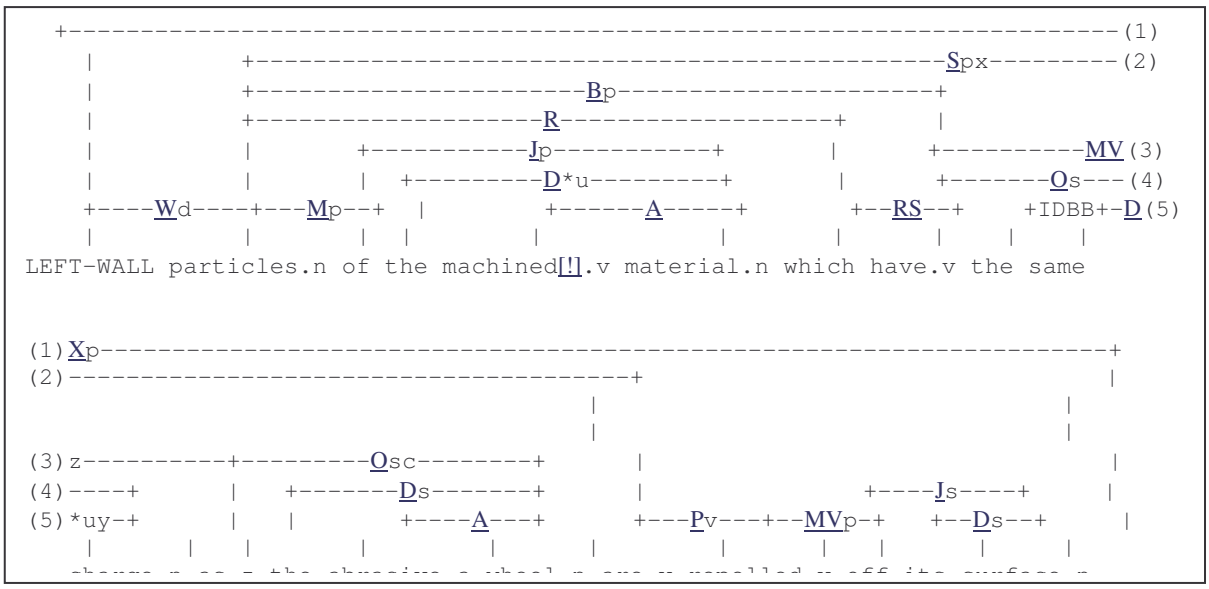


Figure E.4: Analyse de la phrase II par *Link Grammar*

Link Grammar propose plusieurs solutions d'analyse pour chaque phrase. Il en résulte un problème pour la sélection de l'analyse la plus appropriée. Cet outil, dans l'analyse de la phrase I, n'est pas arrivé à identifier le sujet et le complément du verbe *repel*. En revanche, l'analyse n'a présenté aucune faute. Tous les verbes, leurs sujets, et leurs objets sont corrects.

La sortie de cet outil n'est pas utilisable par des agents informatiques. Les liens sont représentés graphiquement par des lignes nommées (cf. figures E.3 et E.4). Cette sortie doit être récupérée et représentée par annotations sous forme d'un texte, d'un tableau ou d'une hiérarchie.

```
MOD_POST(have,wheel)
MOD_POST(repelled,surface)
MOD_PRE(material,machined)
MOD_PRE(charge,same)
MOD_PRE(wheel,abrasive)
MOD_POST(particles,material)
MOD_POST_APPOS(charge,wheel)
DETD(material,the)
DETD(charge,the)
DETD(wheel,the)
DETD(surface,its)
SUBJ_PRE(have,particles)
SUBJ_PRE(are,particles)
NUCL_VLINK_PASSIVE(are,repelled)
OBJ_POST(have,charge)
COORD(and,are)
COORD(and,have)
MAIN(repelled)
MAIN(and)
```

Figure E.5: Analyse de la phrase I par *XIP Pareser*

```
MOD_POST(have,wheel)
MOD_POST(are,surface)
MOD_PRE(material,machined)
MOD_PRE(charge,same)
MOD_PRE(wheel,abrasive)
MOD_POST(particles,material)
MOD_POST_APPOS(charge,wheel)
MOD_POST_SENTENCE_RELATIV(material,have)
DETD(material,the)
DETD(charge,the)
DETD(wheel,the)
DETD(surface,its)
SUBJ_PRE(are,charge)
SUBJ_PRE_RELATIV(have,which)
MAIN(are)
```

Figure E.6: Analyse de la phrase II par *XIP Pareser*

La qualité de l'analyse effectuée par *XIP Pareser* dans la phrase I (cf. figure E.5) et la phrase II (cf. figure E.6) ressemble à celle de *Connexor Machinese Syntax*. Certains composants syntaxiques des phrases sont mis en relation. Le résultat de cette analyse est très clairement présenté et très facile à utiliser pour extraire les triplets. Dans notre système

présenté en chapitre 5, cet outil pourrait être utilisé pour effectuer un post-traitement améliorant la performance de l'extraction.

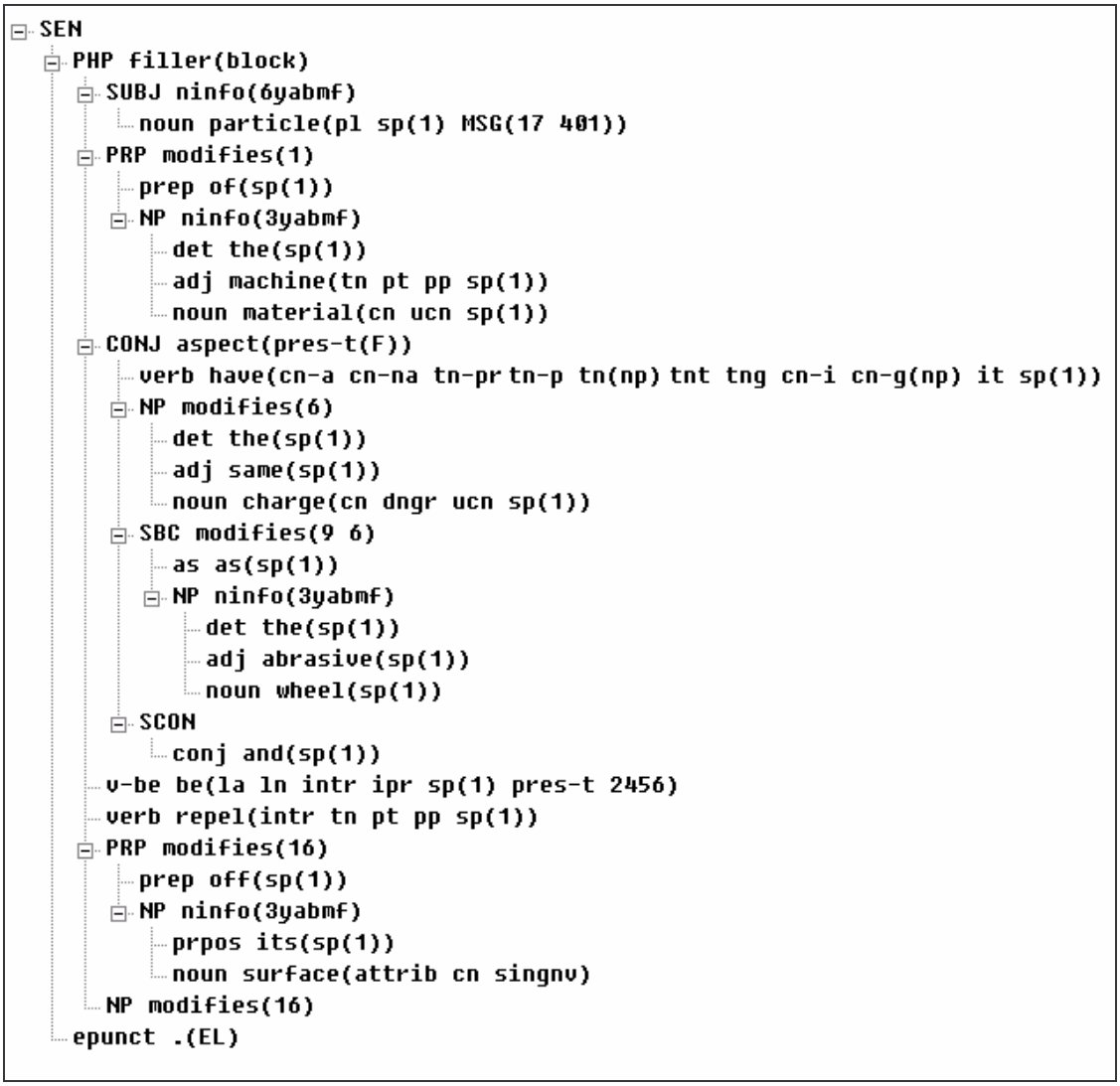


Figure E.7: Analyse de la phrase I par *Proximity Technology Parser*

Dans l'analyse effectuée par *Proximity Technology Parser* (cf. figures E.7 et E.8), les relations sujet-verbe sont très clairement présentées sous forme hiérarchique. Mais les relations verbe-objet ne sont pas fournies dans ce modèle. Un traitement ultérieur est nécessaire pour les extraire.

En conclusion, nous remarquons que ces outils ont tous des performances comparables pour repérer des relations sujet–verbe-objet sauf le dernier, *Proximity Technology Parser*, qui ne fournit pas les relations verbe-objet. Cependant l'extraction des triplets à partir des analyses effectuées par ces analyseurs ne convient pas à notre application pour différentes raisons :

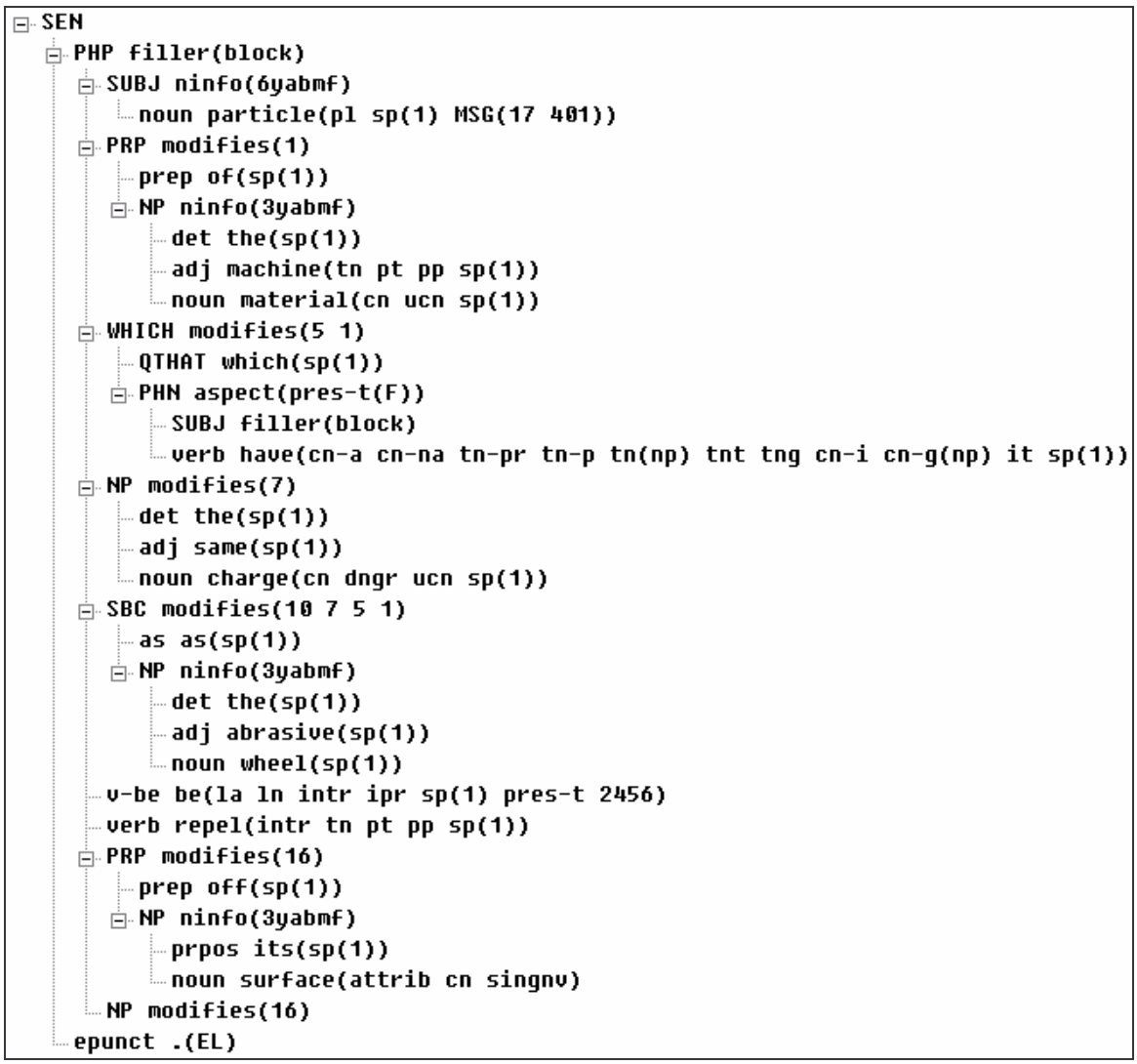


Figure E.8: Analyse de la phrase II par *Proximity Technology Parser*

1. Un post traitement est indispensable afin d'extraire les triplets (sujet, verbe, objet) et les normaliser pour permettre la comparaison avec les triplets des opérateurs d'innovation.

2. Une comparaison d'un triplet (sujet, verbe, objet) extrait par l'analyse avec un triplet (ressource, action, objet) représentant un opérateur d'innovation ne peut pas être faite directement. En effet, le premier type de triplet est syntaxique (apparaît dans une même phrase) et l'autre triplet est sémantique (il correspond à une connaissance du domaine). Par exemple, à partir de l'analyse des phrases I et II, nous pouvons extraire dans le meilleur cas les triplets (sujet, verbe, objet) suivants t_1= (particle, have, charge) et t_2= (, repel, particle). La loi de Coulomb peut être représentée par un triplet T= (charge, repel, objet). La comparaison de T avec t_1 et avec t_2 ne donne

rien, car le sujet du triplet t_1 est inconnu et celui de t_2 est un hyperonyme de *charge* mais l'inverse n'est pas vrai. Donc ni t_1 ni t_2 ne représente un exemple valide de cette loi.

3. L'analyseur ne nous donne pas toutes les annotations nécessaires (les termes associés aux ressources, les objets directs et indirects). Des règles d'extraction, que l'on voudrait ne pas utiliser, sont nécessaires pour repérer ces entités.

Annexe F

Classes lexicographiques des *synsets* de WordNet

Chaque *synset* dans WordNet est associé à un numéro de 00 à 44. Un numéro représente une classe *lexicographique*. Cette classification permet d'organiser les *synsets* et de les séparer en 44 classes. Ces numéros et leurs significations sont présentés dans le tableau suivant.

File Number	Name	Contents
00	adj.all	all adjective clusters
01	adj.pert	relational adjectives (pertainyms)
02	adv.all	all adverbs
03	noun.Tops	unique beginners for nouns
04	noun.act	nouns denoting acts or actions
05	noun.animal	nouns denoting animals
06	noun.artifact	nouns denoting man-made objects
07	noun.attribute	nouns denoting attributes of people and objects
08	noun.body	nouns denoting body parts
09	noun.cognition	nouns denoting cognitive processes and contents
10	noun.communication	nouns denoting communicative processes and contents
11	noun.event	nouns denoting natural events
12	noun.feeling	nouns denoting feelings and emotions
13	noun.food	nouns denoting foods and drinks
14	noun.group	nouns denoting groupings of people or objects
15	noun.location	nouns denoting spatial position
16	noun.motive	nouns denoting goals
17	noun.object	nouns denoting natural objects (not man-made)
18	noun.person	nouns denoting people
19	noun.phenomenon	nouns denoting natural phenomena
20	noun.plant	nouns denoting plants
21	noun.possession	nouns denoting possession and transfer of possession
22	noun.process	nouns denoting natural processes
23	noun.quantity	nouns denoting quantities and units of measure

24	noun.relation	nouns denoting relations between people or things or ideas
25	noun.shape	nouns denoting two and three dimensional shapes
26	noun.state	nouns denoting stable states of affairs
27	noun.substance	nouns denoting substances
28	noun.time	nouns denoting time and temporal relations
29	verb.body	verbs of grooming, dressing and bodily care
30	verb.change	verbs of size, temperature change, intensifying, etc.
31	verb.cognition	verbs of thinking, judging, analyzing, doubting
32	verb.communication	verbs of telling, asking, ordering, singing
33	verb.competition	verbs of fighting, athletic activities
34	verb.consumption	verbs of eating and drinking
35	verb.contact	verbs of touching, hitting, tying, digging
36	verb.creation	verbs of sewing, baking, painting, performing
37	verb.emotion	verbs of feeling
38	verb.motion	verbs of walking, flying, swimming
39	verb.perception	verbs of seeing, hearing, feeling
40	verb.possession	verbs of buying, selling, owning
41	verb.social	verbs of political and social activities and events
42	verb.stative	verbs of being, having, spatial relations
43	verb.weather	verbs of raining, snowing, thawing, thundering
44	adj.ppl	participial adjectives

Annexe G
Base de connaissances utilisée dans les expérimentations

Notre base de connaissances utilisée dans les expérimentations est décrite par le fichier RDF ci-après. Elle est composée de quatre opérateurs d'innovation du type effet (*Effect*). Ces opérateurs sont : loi de Coulomb, fragmentation de liquide par cavitation, pression à partir de l'impact liquide-solide, évaporation de liquide en diminuant la pression de sa vapeur saturée. Des textes et des exemples descriptifs de ces opérateurs sont présentés en langue naturelle dans l'annexe D. Ces opérateurs sont nommés respectivement (par l'attribut *kb:name* de chaque classe RDF) dans la base de connaissances : *Coulomb_law*, *cavitation*, *fragmentation*, *pressure_from_impact* et *vaporization_by_pressure*. Ils ont respectivement les ressources suivantes à améliorer : *attraction force*, *repulsion force*, *fragmentation*, *wave*, *vaporization.* Leurs ressources causes (les ressources associées aux effets par la relation *cause*) sont respectivement : *charge*, *cavity* (trou en gaz ou en vapeur), *impact* et *pressure*. D'autres ressources comme *physical_phenomenon*, *physical_object*, *artefact*, *substance*, *plasma*, *liquide*, *gas* sont ajoutées pour permettre le repérage des objets sur lesquels un opérateur est réalisé. Une relation d'hyperonymie/hyponymie est définie entre les ressources de la base.

- *cavity* est hyponyme de *gas.*
- *plasma*, *liquid*, *gas* sont des hyponymes de *substance*.
- *substance*, *artefact* sont des hyponymes de *physical_object.*
- *impact* est hyponyme de *pressure.*
- *attraction*, *repulsion* sont des hyponymes de *force.*
- *force*, *charge*, *wave*, *pressure* sont des hyponymes de *physical_phenomenon.*

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE rdf:RDF (View Source for full doctype...)>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:kb="http://protege.stanford.edu/kb#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  - <kb:Substance rdf:about="http://protege.stanford.edu/kb#artefact" kb:name="artefact"
      rdfs:label="artefact">
      <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#physical_object" />
    </kb:Substance>
  - <kb:Action rdf:about="http://protege.stanford.edu/kb#attraction" kb:name="attraction"
      rdfs:label="attraction">
      <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#force" />
    </kb:Action>
  - <kb:Effect rdf:about="http://protege.stanford.edu/kb#cavitation_fragmentation"
      kb:name="cavitation_fragmentation" rdfs:label="cavitation_fragmentation">
      <kb:cause rdf:resource="http://protege.stanford.edu/kb#cavity" />
      <kb:improve rdf:resource="http://protege.stanford.edu/kb#fragmentation" />
    </kb:Effect>
  - <kb:Substance rdf:about="http://protege.stanford.edu/kb#cavity" kb:name="cavity"
      rdfs:label="cavity">
      <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#gas" />
    </kb:Substance>
  - <kb:Action rdf:about="http://protege.stanford.edu/kb#charge" kb:name="charge"
      rdfs:label="charge">
      <kb:hypernym
        rdf:resource="http://protege.stanford.edu/kb#physical_phenomenon" />
    </kb:Action>
  - <kb:Effect rdf:about="http://protege.stanford.edu/kb#coulomb_law1"
      kb:name="coulomb_law" rdfs:label="coulomb_law1">
      <kb:cause rdf:resource="http://protege.stanford.edu/kb#charge" />
      <kb:improve rdf:resource="http://protege.stanford.edu/kb#repulsion" />
    </kb:Effect>
  - <kb:Effect rdf:about="http://protege.stanford.edu/kb#coulomb_law2"
      kb:name="coulomb_law" rdfs:label="coulomb_law2">
      <kb:improve rdf:resource="http://protege.stanford.edu/kb#attraction" />
      <kb:cause rdf:resource="http://protege.stanford.edu/kb#charge" />
    </kb:Effect>
  - <kb:Energy rdf:about="http://protege.stanford.edu/kb#force" kb:name="force"
      rdfs:label="force">
      <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#attraction" />
      <kb:hypernym
        rdf:resource="http://protege.stanford.edu/kb#physical_phenomenon" />
      <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#repulsion" />
    </kb:Energy>
  - <kb:State rdf:about="http://protege.stanford.edu/kb#gas" kb:name="gas"
      rdfs:label="gas">
      <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#cavity" />
      <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#substance" />
    </kb:State>
  - <kb:Action rdf:about="http://protege.stanford.edu/kb#impact" kb:name="impact"
      rdfs:label="impact">
      <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#pressure" />
    </kb:Action>
  - <kb:State rdf:about="http://protege.stanford.edu/kb#liquid" kb:name="liquid"
      rdfs:label="liquid">
      <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#substance" />
    </kb:State>
  - <kb:Substance rdf:about="http://protege.stanford.edu/kb#physical_object"
```

```
kb:name="physical_object" rdfs:label="physical_object">
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#artefact" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#substance" />
  </kb:Substance>
- <kb:Energy rdf:about="http://protege.stanford.edu/kb#physical_phenomenon"
    kb:name="physical_phenomenon" rdfs:label="physical_phenomenon">
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#charge" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#force" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#pressure" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#wave" />
  </kb:Energy>
- <kb:State rdf:about="http://protege.stanford.edu/kb#plasma" kb:name="plasma"
    rdfs:label="plasma">
    <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#substance" />
  </kb:State>
- <kb:Action rdf:about="http://protege.stanford.edu/kb#pressure" kb:name="pressure"
    rdfs:label="pressure">
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#impact" />
    <kb:hypernym
      rdf:resource="http://protege.stanford.edu/kb#physical_phenomenon" />
  </kb:Action>
- <kb:Effect rdf:about="http://protege.stanford.edu/kb#pressure_from_impact"
    kb:name="pressure_from_impact" rdfs:label="pressure_from_impact">
    <kb:cause rdf:resource="http://protege.stanford.edu/kb#impact" />
    <kb:improve rdf:resource="http://protege.stanford.edu/kb#wave" />
  </kb:Effect>
- <kb:Action rdf:about="http://protege.stanford.edu/kb#repulsion" kb:name="repulsion"
    rdfs:label="repulsion">
    <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#force" />
  </kb:Action>
  <kb:Action rdf:about="http://protege.stanford.edu/kb#fragmentation"
    kb:name="fragmentation" rdfs:label="fragmentation" />
- <kb:State rdf:about="http://protege.stanford.edu/kb#solid" kb:name="solid"
    rdfs:label="solid">
    <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#substance" />
  </kb:State>
- <kb:Substance rdf:about="http://protege.stanford.edu/kb#substance"
    kb:name="substance" rdfs:label="substance">
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#gas" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#liquid" />
    <kb:hypernym rdf:resource="http://protege.stanford.edu/kb#physical_object" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#plasma" />
    <kb:hyponym rdf:resource="http://protege.stanford.edu/kb#solid" />
  </kb:Substance>
  <kb:Action rdf:about="http://protege.stanford.edu/kb#vaporization"
    kb:name="vaporization" rdfs:label="vaporization" />
- <kb:Effect rdf:about="http://protege.stanford.edu/kb#vaporization_by_pressure"
    kb:name="vaporization_by_pressure" rdfs:label="vaporization_by_pressure">
    <kb:cause rdf:resource="http://protege.stanford.edu/kb#pressure" />
    <kb:improve rdf:resource="http://protege.stanford.edu/kb#vaporization" />
  </kb:Effect>
- <kb:Action rdf:about="http://protege.stanford.edu/kb#wave" kb:name="wave"
    rdfs:label="wave">
    <kb:hypernym
      rdf:resource="http://protege.stanford.edu/kb#physical_phenomenon" />
  </kb:Action>
</rdf:RDF>
```

Annexe H

Collection de documents utilisés pour les expérimentations et résultats de l'évaluation

Dans cette annexe nous présentons un tableau descriptif de la collection de documents utilisée, le tableau explicitant l'annotation manuelle, la table du mapping des ressources à des *synsets* dans WordNet et le tableau des résultats pour l'évaluation du système.

H.I Tableau descriptif de la collection

Ce tableau présente dans la colonne « *lien vers le texte* » les 63 documents de la collection d'expérimentation. La colonne « *texte* » est utilisée comme référence à ces documents dans les différents tableaux de cette annexe. La colonne *taille* représente en kilooctets la taille de chaque texte.

Texte	**Lien vers le texte**	**Taille**
texte1	http://www.freepatentsonline.com/EP0208647.html	9
texte2	http://www.patentstorm.us/patents/4522843/fulltext.html	13
texte3	http://www.patentmonkey.com/PM/patentid/5817374.aspx	17
texte4	http://www.wipo.int/pctdb/en/wo.jsp?wo=1986003993	5
texte5	http://reporter-archive.mcgill.ca/Rep/r3013/marchessault.html	9
texte6	http://www.utulsa.edu/news/article.asp?Key=905	6
texte7	Van de Graaff's Invention : http://mig.rssi.ru/.../Svandgrf.htm	6
texte8	http://www.answers.com/topic/atmospheric-electricity-2	5
texte9	http://www.chargesyndrome.ca/	2
texte10	http://www.liquorsnob.com/.../liquid_core_liquid_charge_energy_drink_review.php	1
texte11	http://www.freepatentsonline.com/6534129.html	33
texte12	abrasive wheel : goldefire innovetor	1
texte13	Lightning : http://mig.rssi.ru/.../Svandgrf.htm	1
texte14	http://www.purchon.com/chemistry/ions.htm	6
texte15	http://www.sciencemadesimple.com/static.html	4
texte16	http://physics.bu.edu/~duffy/PY106/Charge.html	12

texte17	page5 : http://www.scribd.com/doc/3886448/new-developments-in-theoretical-physics	3
texte18	http://www.exploratorium.edu/snacks/charge_carry/index.html	7
texte19	http://www.patentstorm.us/patents/6118329/fulltext.html	22
texte20	http://www.historyoftheuniverse.com/charge.html	2
texte21	http://www.glenbrook.k12.il.us/gbssci/phys/mmedia/estatics/itsn.html	4
texte22	http://www.glenbrook.k12.il.us/GBSSCI/PHYS/mmedia/estatics/esn.html	4
texte23	Abstract: http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/7959/22002/01022530.pdf?arnumber=1022530	1
texte24	Electroscope : http://www.thebakken.org/electricity/science-of-static.html	1
texte25	http://www.srh.noaa.gov/jetstream//lightning/positive.htm	3
texte26	http://www.lightningsafety.noaa.gov/media/video_text_science.htm	6
texte27	http://www.eskimo.com/~billb/emotor/belt.html	10
texte28	http://www.freepatentsonline.com/5295083.html	4
texte29	http://www.geocities.com/davidmdelaney/geyser/Sorensens-bubble-pump.html	2
texte30	http://www.freepatentsonline.com/7261144.html	37
texte31	http://www.patentstorm.us/patents/6854831/fulltext.html	123
texte32	http://www.freepatentsonline.com/6737393.html	6
texte33	http://www.freepatentsonline.com/5947784.html	24
texte34	http://hyperphysics.phy-astr.gsu.edu/Hbase/Kinetic/vappre.html	1
texte35	http://www.freepatentsonline.com/6582552.html	13
texte36	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V3H-4FR8PST2&_user=636532&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_version=1&_urlVersion=0&_userid=636532&md5=89d01849b4f8a660295686ec65dab019	4
texte37	http://www.bronkhorst.com/en/products/vapour_delivery_systems/	1
texte38	http://urila.tripod.com/colligative.htm	6
texte39	http://www.freepatentsonline.com/4704189.html	6
texte40	http://www.freepatentsonline.com/6830654.html	11
texte41	http://www.freepatentsonline.com/6837969.html	13
texte42	http://www.freepatentsonline.com/6656327.html	19
texte43	http://www.freepatentsonline.com/5795446.html	19
texte44	http://www.sciencemag.org/cgi/content/abstract/295/5558/1261	2
texte45	http://web.wits.ac.za/Academic/EBE/MechEng/Research/ResearchUnits/FRU/Research+Areas/LiquidShockWaves.htm	2
texte46	http://www.patentstorm.us/patents/5394786/fulltext.html	91
texte47	http://www.springerlink.com/content/ggljyn0eq70w4pdc/	1
texte48	Abstract: http://journals.pepublishing.com/content/yg00h87870210465/	1
texte49	Abstract: http://cat.inist.fr/?aModele=afficheN&cpsidt=10709624	2
texte50	http://www.iit.edu/~smile/ph8808.html	3
texte51	http://arxiv.org/abs/cond-mat/0109349	2
texte52	http://www.iit.edu/~smile/ph8904.html	4

texte53	http://www.wipo.int/pctdb/en/wo.jsp?wo=1998046684	2
texte54	http://chronicle.uchicago.edu/070426/light.shtml	5
texte55	http://www.fuji-electric.com/uk/catalogue/catalogue.php4?gamme=11	2
texte56	http://www.intmath.com/Applications-integration/10_Force-Liquid.php	2
texte57	http://pubs.nrc-cnrc.gc.ca/cjche/ch82871-5.html	1
texte58	http://www.freepatentsonline.com/6186963.html	16
texte59	fragment rock : goldefire innovetor	1
texte60	http://www.freepatentsonline.com/6647910.html	13
texte61	http://www.freepatentsonline.com/5423917.html	28
texte62	http://www.freepatentsonline.com/6534129.html	10
texte63	http://www.freepatentsonline.com/6666834.html	15

H.II Tableau de l'annotation manuelle

Ce tableau est élaboré pour mémoriser les annotations manuelles afin de les comparer avec les annotations automatiques pour l'évaluation du système. Ces annotations définissent les termes représentant des ressources pertinentes et les opérateurs pour lesquels le document constitue un exemple. Dans ce tableau les opérateurs sont représentés par les colonnes Ca, Co, Va, Wa. Ces colonnes correspondent respectivement aux opérateurs suivants : *fragmentation de liquide par cavitation*, *loi de coulomb*, *évaporation de liquide en diminuant la pression de sa vapeur saturée*, *pression à partir de l'impact liquide-solide*, Ces opérateurs ont été présentés dans l'annexe D.

Texte	**Termes annotés comme ressource objet**	**Exemple annoté**			
		Ca	**Co**	**Va**	**Wa**
texte1	support, material, surface	0	1	0	0
texte2	paper, acide, book, material, image	0	1	0	0
texte3	material, particle, surface, mask	0	1	0	0
texte4	stream, material, product	0	1	0	0
texte5	surface, image, paper, powder	0	1	0	0
texte6		0	0	0	0
texte7		0	0	0	0
texte8		0	0	0	0
texte9		0	0	0	0
texte10		0	0	0	0
texte11	material, substrate, liquid, source, bead, placement	0	1	0	0
texte12	wheel, surface, particle, material	0	1	0	0
texte13	fragment	0	1	0	0
texte14		0	0	0	0
texte15	balloon, wall, hair,	0	1	0	0
texte16	pen, stream	0	1	0	0
texte17		0	0	0	0
texte18	nail, body, jar, pan, foil, plate, cloth, styrofoam	0	1	0	0

texte19		0	0	0	0
texte20	balloon, ceiling, particle, jumper	0	1	0	0
texte21	balloon, sphere, conductor, induction	0	1	0	0
texte22	electroscope, balloon, plate	0	1	0	0
texte23		0	0	0	0
texte24	electroscope, ball, object, electricity	0	1	0	0
texte25		0	0	0	0
texte26		0	0	0	0
texte27	roller, metal, tip, wind, surface, needle, plasma, air, belt, teeth, machine	0	1	0	0
texte28		0	0	0	0
texte29	pump, vapour, water	1	0	0	0
texte30	pump, microchannel, assembly, heat, fluid, chip	1	0	0	0
texte31	Discharge	1	0	0	1
texte32		0	0	0	0
texte33	fluid, water, air	0	0	0	1
texte34		0	0	0	0
texte35	tank, fluid, exchanger, heat, supply, device, circulation.	0	0	1	0
texte36		0	0	0	0
texte37		0	0	0	0
texte38	molecule, solute	0	0	1	0
texte39	medium, air, solution	0	0	1	0
texte40	gas, water	0	0	1	0
texte41		0	0	0	0
texte42	shell, vessel, water, steam	0	0	1	0
texte43	exchanger, bubble, equipment, steam	1	0	0	0
texte44		0	0	0	0
texte45	metal, tube	0	0	0	1
texte46	assembly, medium, liquid, material	1	0	0	1
texte47	fluorocarbon, water	0	0	0	1
texte48	deformation, plate, tube	0	0	0	1
texte49	electromagnetic source, acceleration, disk, water, jet	1	0	0	1
texte50		0	0	0	0
texte51	helium,	1	0	0	1
texte52		0	0	0	0
texte53	ink, printing,	0	0	0	0
texte54		0	0	0	0
texte55		0	0	0	0
texte56		0	0	0	0
texte57		0	0	0	0
texte58	tissue, body, spark, electrode, discharge	0	0	0	1
texte59	rock, material, divice, water, disk, pressure	0	0	0	1
texte60	liquid, air, hull, vehicule, hull	0	0	0	1
texte61	tube, wave	0	0	1	1
texte62		0	0	0	0
texte63	electrode, voltage, device	0	1	1	1

H.III Mapping des termes extraits à WordNet

resouce	offset
physical_object	16236
physical_phenomenon	10681095
substance	17572
pressure	105390
repulsion	903958
attraction	4493818
shock_wave	6895233
charge	8693651
charge	8809850
attraction	10688069
repulsion	10688453
wave	06900919
force	10719108
pressure	10753257
liquid	14090852
liquid	14091166
cavitation	13123142
segmentation	00380410
gas	14030133
artefact	00019244
impact	01107002
gas	14201971
vaporization	12807580
plasma	14140840

Les termes annotés dans la collection et présentés dans le tableau précédent sont mappés aux ressources suivantes : *physical_phenomenon*, *physical_object*, *artefact*, *substance*, *plasma*, *liquide*, *gas*. Ce mapping est installé par le moyen de la table relationnelle *Map1Expert* présentée ci-après. Dans cette table, les ressources sont mappées aux *synsets* correspondant à leurs termes dans WordNet. Des ressources (i.e. charge, pressure, attraction, force) existent déjà dans *Map1Expert*. Elles constituent les ressources spécifiques aux opérateurs définis dans la base de connaissances. Elles sont mappées à leurs *synsets* lorsqu'elles sont définies dans la base de connaissances. Pour des raisons de simplicité, les termes correspondant à ces ressources spécifiques ne sont pas reproduits dans le tableau de l'annotation manuelle de la section précédente.

Par ce mapping des termes comme *ballon*, *electroscope*, *pump*, *roller* sont mappés implicitement à la ressource *artefact*. Des termes comme :

- *discharge*, *induction*, *electricity*, *voltage* sont mappés à la ressource *physical_phenomenon* ;
- *water*, *ink* sont mappés à la ressource *liquid* ;
- *helium*, *air* sont mappés à la ressource *gas ;*
- *material*, *body*, *rock*, *molecule* sont mappés à la ressource *physical_object ;*
- *fluid*, *solute*, *styrofoam* sont *mappés* à la ressource *substance.*

Dans notre base de connaissances, les termes ajoutés dans ce mapping sont associés par la relation d'hyperonymie/hyponymie.

H.IV Tableau de l'évaluation du système

Présenté ci-après, ce tableau est composé de deux parties. Chaque partie est composée de trois colonnes $N_{correctes}$, $N_{réponses}$ et $N_{clés}$. Elles représentent respectivement le nombre des annotations correctes, le nombre des annotations automatiques et le nombre des annotations manuelles. Dans la partie gauche, ces annotations se rapportent aux ressources (i.e. aux associations terme-ressource) et dans la partie droite elles se rapportent aux exemples (i.e. aux associations document-opérateur). Les résultats reportés dans ce tableau nous permettent de calculer la précision et le rappel pour la collection de documents.

***Note* 1 :** $N_{correctes}$ des opérateurs peut être égal à zéro dans deux cas : dans le cas où le système extrait du document des exemples et aucun exemple n'est annoté manuellement et dans le cas où le document contient des exemples mais le système n'en repère aucun.

***Note 2* :** l'annotation des termes par leurs ressources et l'annotation des documents par leurs opérateurs sont traitées comme des tâches séparées. Ainsi, un document qui ne représente pas un exemple d'un opérateur peut contenir des termes représentant des ressources pertinentes. $N_{correctes}$ des exemples peut donc être égal à zéro et $N_{correctes}$ des ressources a toujours une valeur différente de zéro.

***Note 3* :** $N_{clés}$ des exemples, ne peut pas être égal à zéro, car le document est au pire annoté comme impertinent et donc $N_{clés}$=1. Sinon $N_{clés}$ est le nombre des opérateurs dont le document constitue un exemple dans le tabelau de l'annotation manuelle.

Texte	**Ressource**			**Exemple**		
	$N_{correctes}$	$N_{réponses}$	$N_{clés}$	$N_{correctes}$	$N_{réponses}$	$N_{clés}$
texte1	4	5	4	1	1	1
texte2	22	26	25	1	1	1
texte3	11	26	11	1	1	1
texte4	5	5	7	1	1	1
texte5	14	16	17	1	1	1
texte6	0	0	0	1	1	1
texte7	0	0	0	1	1	1
texte8	0	0	0	1	1	1
texte9	0	0	0	1	1	1
texte10	0	0	0	1	1	1
texte11	0	0	6	0	0	1
texte12	6	6	6	1	1	1
texte13	10	13	10	1	1	1
texte14	6	8	6	1	1	1
texte15	7	13	9	1	1	1
texte16	12	14	14	1	2	1
texte17	0	0	0	1	1	1
texte18	0	0	8	1	1	1
texte19	0	0	0	1	1	1
texte20	9	9	10	1	1	1
texte21	10	11	10	1	3	1
texte22	12	12	12	1	1	1
texte23	0	0	0	1	1	1
texte24	0	0	4	0	0	1
texte25	0	0	0	1	1	1
texte26	0	3	0	1	2	1
texte27	16	17	24	1	1	1
texte28	0	0	0	1	1	1
texte29	24	29	24	1	2	1
texte30	40	60	40	1	3	1
texte31	15	31	15	2	3	2
texte32	0	0	0	1	1	1
texte33	0	0	3	0	0	1
texte34	8	11	8	1	2	1
texte35	16	27	19	1	1	1
texte36	12	15	12	1	1	1
texte37	0	0	0	1	1	1
texte38	17	27	17	1	1	1
texte39	13	14	15	1	2	1
texte40	22	33	22	1	2	1
texte41	11	14	11	1	2	1
texte42	26	39	26	1	1	1
texte43	30	49	30	1	2	1
texte44	0	2	0	1	1	1

texte45	15	15	15	1	1	1
texte46	37	41	37	0	0	2
texte47	5	6	6	1	1	1
texte48	5	5	8	1	1	1
texte49	12	12	14	2	2	2
texte50	0	0	0	1	1	1
texte51	4	4	4	0	0	2
texte52	8	10	8	1	2	1
texte53	0	0	0	1	1	1
texte54	0	0	0	1	1	1
texte55	0	0	0	1	1	1
texte56	0	0	0	1	1	1
texte57	0	0	0	1	1	1
texte58	9	14	12	1	1	1
texte59	0	0	6	0	0	1
texte60	0	0	5	0	0	1
texte61	23	38	23	2	2	2
texte62	0	0	0	1	1	1
texte63	19	22	21	3	3	3
somme	515	702	584	61	75	70
score	**Précision (P)**	**Rappel (R)**	**P&R**	**Précision (P)**	**Rappel (R)**	**P&R**
	73,36%	88,18%	80,09%	81,33%	87,14%	84,14%

TITRE : Alimentation automatique d'une base de connaissances à partir de textes en langue naturelle

RESUME : Dans ce travail nous nous sommes intéressés à l'alimentation automatique d'une base de connaissances pour l'aide à l'innovation. Ce processus s'appuie sur une ontologie du domaine. La base de connaissances est organisée autour des opérateurs d'innovation. Cette base est initialisée par un expert qui doit définir les opérateurs concernés et les ressources associées. Le système d'alimentation automatique permet alors l'enrichissement de cette base par des exemples de résolution de problèmes d'innovation à partir de textes en langue naturelle. Ce système met en œuvre une nouvelle approche pour l'extraction automatique d'informations. Cette approche n'est pas spécifique à l'innovation et peut être adaptée à d'autres problèmes d'extraction d'informations dans d'autres domaines.

MOTS-CLES : base de connaissances, enrichissement automatique extraction d'information, ontologie, système de question-réponse, règle d'extraction, texte en langue naturelle, innovation, TRIZ, innovation operateur, innovation ressource

TITLE: Automatic feeding of a knowledge base starting from natural language texts

ABSTRACT: In this work, we were interested in automatic feeding of a knowledge base for innovation aid. This process relies on domain ontology. The knowledge base is organized around innovation operators. It is initialized by an expert who must define the operators and their associated innovation resources. Then the automatic feeding system allows the enrichment of this base by examples of inventive problem-solving from natural language texts. This system implements a new information extraction approach. This approach is not specific to the innovation domain and can be adapted to other problems for extracting information in other domains.

KEYWORDS: knowledge base, automatic feeding, information extraction, ontology**,** question answer system, extraction rule, natural language text, innovation, TRIZ, innovation operator, innovation resource